AfIA

Association française
pour l'Intelligence Artificielle

CNIA

29e Conférence Nationale en Intelligence Artificielle

PFIA 2026



Table des matières

Florence Dupin de Saint-Cyr
Éditorial 6

Comité de programme 7

Présentation invitée CNIA 2026 8

Nahla Ben Amor
Biais dans les données : une approche agentique pour orchestrer l'équité 9

SESSION 1 : Incertitude et classification présidée par Henri Prade 11

François Thievon, Arthur Hoarau
Et si les classifieurs profonds oubliaient l'inconnu ? Auto-encodeurs et détection hors distribution 12

Anna Pappa, Pierre Pomeret-Coquot, Francis Faux, Revekka Kyriakoglou
Annotation linguistique sous incertitude structurelle, vers un modèle ensembliste basé sur les fonctions de croyance 21

SESSION 2 : Apprentissage automatique et regroupement présidée par Florence Dupin de Saint-Cyr 31

Myriam Bounhas, Henri Prade
Des proportions logiques à un nouvel algorithme de clustering 32

Jean-Charles Lamirel, Christophe Malaterre, Francis Lareau, Thibault Prouteau
Vieux mais pas obsolètes. Bag-of-Words vs Embeddings en modélisation thématique 43

Rosalie Nkounghawe, Paulin Melatagia, Isabelle Pieretti, Carlène Trévenec, Mathieu Roche
Processus d'identification automatique d'articles de presse dédiés à la veille syndromique 45

Hamid Ahaggach, Lylia Abrouk,Éric Lebon
Prédiction du délai de vente et optimisation des stocks par apprentissage automatique 54

SESSION 3 : IA, Éthique et Enseignement présidée par Anna Pappa 58

Edouard Nadaud, Lionel Prévost, Bénédicte Le Grand, Bastien Tarot
Indicateurs émotionnels et posturaux pour l'intervention pédagogique précoce, modéliser la difficulté comme une déviation de la réussite individuelle de l'étudiant. 59

Robert Viseur
La troisième voie des « Code LLM » ouverts et éthiques, une stratégie de différenciation durable face aux modèles extractivistes ? 70

Thierry Bénard
Une balance à levier comme neurone. Un système mécanique pédagogique pour l'enseignement des principes fondamentaux de l'IA 74

SESSION 4 : Explicabilité et apprentissage de structure présidée parIsmaïl Baaj 79

Fabien Lionti, Nicolas Gutowski, Philippe Martinet, Sébastien Aubin
Apprentissage de dynamiques latentes. Stabilisation par structure hamiltonienne 80

Jordan Levy, Paul Saves, Moncef Garouani, Nicolas Verstaevel, Benoit Gaudou
Caractérisation de la complémentarité des détecteurs d'anomalies par l'analyse des contributions SHAP 96

Ali Assi, Nour Elislem Karabadji, Mohamed Elati, Wajdi Dhifli
Apprentissage Multi Échelle Global Local de Représentations de Noeuds par Marches Aléatoires

et Optimisation Guidée par Points de Repère 105

Bounia Louenas
Explications Probabilistes Préférées Pareto-Optimales. Équilibrer Contraintes Cognitives et Préférences Utilisateur 107

SESSION 5 : SAT, ASP présidée parFrédéric Maris 119

Meriam Ayari, Sami Cherif, Felip Manyà
Sur le calcul des max réfutations minimales 120

Carl Berton, Sami Cherif, Claire Delaplace
Satisfiabilité pour le problème du décodage par syndrome à poids élevé 132

Asma Khoualdia, Sami Cherif, Stéphane Devismes, Léo Robert
Une approche SAT pour l'analyse de l'autostabilisation de l'unisson asynchrone 143

Igor Stéphan, Martin DIEGUEZ
Extraction des politiques pour la programmation par ensembles réponses quantifiée 154

François Olivier, Zied Bouraoui
Vers un système de raisonnement neurosymbolique ancré dans des représentations schématiques 156

SESSION 6 : Paradigme multi-agents présidée parElsy Kaddoum 160

Alaa Daoud, Hiba Alqasir, Maxime Gueriau, Corentin Bunel
Fusion de détecteurs d'objets par coordination multi-agent et a priori causaux 161

Laura Trujillo, Elsy Kaddoum, Gauthier Picard
Approche par optimisation sous contraintes distribuée "Divide-and-Coordinate" pour le transport urbain multimodal 165

SESSION 7 : IA et Humains présidée par Nahla Ben Amor 175

Hidri Aya, Inès Saad, Manel Ben Sassi, Henda Ben Ghezala
Approche hybride centrée sur l'humain pour l'évaluation explicable de la qualité de présentation des tweets organisationnels 176

Jules Leguy, Nicolas Soulié, Felipe Torres Figueroa, Andon Tchechmedjiev, Jacky Montmain, Sébastien Harispe
Effets d'ancrage liés à la pression temporelle et à la difficulté des tâches dans un contexte de décision humaine assistée par IA 187

Robert Viseur
Évaluer la sobriété d'un service numérique. Proposition d'une grille d'évaluation multicritère adaptée au « fuzzy front end » 199

Safa Yahi
Proposition d'un cadre argumentatif pour la Communication Non Violente 209

SESSION 8 : Apprentissage et classification présidée parAstrid Klipfelr 213

Malek Adouani, Zaineb Chelly Dagdia
Génération de Données Synthétiques Équitables et Préservant la Vie Privée via Autoencodeur Variationnel Basé sur le Clustering et Réseaux Antagonistes Génératifs 214

Juliette Van Poulle, Alissa Chahine
Étude comparative de méthodes de vision par ordinateur pour la détection de near-duplicates d'images de coraux 216

Farida Nchare, Norbert Tsopze, Corina Iovan
The attention illusion. Overcoming the localization-prediction discrepancy in biomass ViT-based regression 224

Martin C. Cooper, Imane Bousdira
Équité des classifieurs en présence de contraintes entre les attributs 231

Samuel Stocksieker, Denys Pommeret
SHGR, A Generalized Maximal Correlation Coefficient 236

SESSION 9 : Ontologies et logiques de description présidée parNathalie Aussenac 240

Léa Charbonnier, Franco Giustozzi, Julien Saunier, Cecilia Zanni-Merk
A Novel Concept Induction Approach for Explainable Quality 4.0 241

Pierre Lafourcade, Meghyn Bienvenu, Diego Figueira
Mesures de responsabilité à faible complexité pour requêtes en présence d'ontologies 244

SESSION 10 : IA générative, LLM et prompts présidée par Safa Yahi 246

Hamza Bendali, Oussema Ben Henni, Youssef Douzi, Abokor Mahammad Mousse, Frédéric Flouvat, Safa Yahi
Revue des travaux récents sur l'usage des LLMs en cybersécurité offensive 247

Christophe Denis
Le prompt comme opérateur d'action. Fabrique sémiotique hybride et interaction minimale avec les grands modèles de langage 257

Arnaud Deleruyelle, Steve Bellart, Jérôme Mollier-Pierret
JK-Means. LLM-as-a-Judge au Service du Clustering Intentionnel 268

Stergos Afantenos, Leonardo Cortez-Bernades, Henri Prade
Evaluer la créativité analogique. Une expérimentation 279

Éditorial

29e Conférence Nationale en Intelligence Artificielle

La 29e Conférence Nationale en Intelligence Artificielle (CNIA), organisée sous l'égide du Conseil d'Administration de l'AFIA, rassemble l'ensemble de la communauté de recherche en IA. Cet événement phare vise à présenter les dernières avancées du domaine, tout en encourageant les échanges entre ses différentes sous-disciplines et le rayonnement interdisciplinaire. À l'ère de la révolution sociétale portée par l'IA générative et face au succès croissant des outils intelligents, l'IA se trouve aujourd'hui au coeur de nombreux développements. L'objectif de CNIA est d'aborder à la fois les problématiques de recherche, les enjeux technologiques et les enjeux sociétaux liés à l'utilisation de l'IA.

CNIA a eu lieu du 29 Juin au 1^{er} juillet 2026 dans le cadre de la Plate-Forme Intelligence Artificielle (PFIA). Trois types d'articles y ont été présentés :

- Contributions originales : présentation de travaux de recherches et de résultats en lien avec les thématiques de CNIA sous la forme d'article court ou long (4 ou 10 pages références non comprises respectivement).
- Panorama ou synthèse d'un domaine en lien avec CNIA, qui ne soit ni trop étroit ni trop large ; exposé des travaux principaux, des problématiques, verrous scientifiques et technologiques. Taille : 10 pages maximum (références non comprises)
- Articles prospectifs présentant des idées, visions, défis qui incitent la communauté à poursuivre de nouvelles voies de recherche (nouveaux problèmes, nouveaux domaines d'application, nouvelles méthodologies). Les auteurs devaient pouvoir convaincre l'audience que le sujet est intéressant et prometteur, et devaient le relier à l'état de l'art. Taille : 4 pages maximum (références non comprises)
- Contributions déjà publiées : des articles déjà publiés dans des conférences ou revues internationales mais inédits en français. Il fallait soumettre l'article d'origine dans sa langue et taille d'origine et la version pour les actes de CNIA : un résumé de 2 à 4 pages obligatoirement en français (références non comprises).

Les thèmes de CNIA 2026 sont les suivants :

1. Résolution algorithmique de problèmes
   (a) Algorithmes génétiques / évolutionnaires / métaheuristiques
   (b) Choix social computationnel
   (c) Complexité des algorithmes ou des problèmes
   (d) Paradigme multi-agents et émergence, résolution distribuée
   (e) Recherche heuristique, résolution de problèmes, jeux
   (f) SAT, Contraintes, Plannification, Optimisation combinatoire
2. Représentation et raisonnement sur des croyances
   (a) Argumentation
   (b) Décision individuelle / collective (préférences)
   (c) Graphe (graphes conceptuels, graphe de connaissances, ontologies)
   (d) Incertitude
   (e) Logique
   (f) Raisonnement sur le changement (révision, mise-à-jour, actions, causalité)
3. Apprentissage automatique et fouille de données
   (a) Apprentissage par renforcement, Active learning
   (b) Apprentissage profond, CNN, GAN
   (c) Classification, recommandation
   (d) Explicabilité, apprentissage de structure
   (e) Extraction, ingéniérie des connaissances
   (f) IA générative, LLM, Ingéniérie du prompt
4. IA et société
   (a) IA et Droit, IA et Éthique (et sécurité, vie privée)
   (b) IA et Enseignement, EIAH
   (c) IA et Humain (cognition, interaction humain-agent)
   (d) IA et Langage (NLP, linguistique)
   (e) IA et Robotique
   (f) IA et X (santé, environnement, énergie, transport, défense, agriculture, matériaux, ...)

L'édition 2026 a attiré 50 intentions de soumissions (sous la forme de titre et résumé), qui ont été concrétisées par 40 articles soumis. Chaque article a été relu par 2 relecteurs et évalué sur 4 critères : Originalité, Clarté, Correction/Méthodologie, Importance/Impact. Au final 35 articles ont été acceptés pour présentation et publication dans ces actes.

Merci à tous ceux qui ont contribué à la réussite de CNIA'2026.

Florence Dupin de Saint-Cyr

Comité de programme

Présidence

— Florence Dupin De Saint-Cyr - IRIT, Université de Toulouse.

Membres

— Leila Amgoud IRIT, CNRS, Université de Toulouse ;
— Ismaïl Baaj LEMMA, Université Paris Panthéon Assas ;
— Salem Benferhat CRIL, Université d'Artois ;
— Pierre Bisquert IATE INRAE, Montpellier ;
— Isabelle Bloch ENST - CNRS LTCI ;
— Corentin Boidot IRENav, École Navale Brest ;
— Olivier Boissier LIMOS, Mines Saint-Etienne ;
— Anne-Gwenn Bosser Lab-STICC, École Nationale d'Ingénieurs de Brest ;
— Celia da Costa I3S, Université Nice Sophia Antipolis ;
— Louise Dupuis de Tarlé LAMSADE, Université Paris-Dauphine ;
— Béatrice Fuchs LIRIS, Université de Lyon ;
— Jean-Gabriel Ganascia LIP6, Université Pierre et Marie Curie ;
— Andreas Herzig IRIT, CNRS Université de Toulouse ;
— Emiliano Lorini IRIT, CNRS Université de Toulouse ;
— Nassim Mokhtari Lab-STICC, École Nationale d'Ingénieurs de Brest ;
— Véronique Moriceau IRIT, Université de Toulouse ;
— Anna Pappa LIASD, Université Paris 8 ;
— Élise Perrotin University of Bergen, Norway ;
— Gauthier Picard DTIS, ONERA, Toulouse ;
— Stephanie Roussel ONERA, Toulouse ;
— Céline Rouveirol LIPN, Universit Paris 13 ;
— Laurent Vercouter LITIS lab, INSA de Rouen ;
— Srdjan Vesic CRIL, Université d'Artois ;
— Safa Yahi LIS, Université Aix-Marseille ;
— Bruno Zanuttini GREYC, Université de Caen.

Présentation invitée CNIA 2026

Exposé invité CNIA 2026
Nahla Ben Amor

Professeur d'informatique à l'Université de Tunis

LARODEC*

Biais dans les données : une approche agentique pour orchestrer l'équité

Les modèles d'apprentissage automatique sont aujourd'hui utilisés dans des domaines sensibles comme le crédit, le recrutement, la justice ou les services publics. Ils apprennent à partir de données historiques qui portent la trace de décisions passées, et peuvent donc reproduire les biais présents dans ces données. Le débiaisage adversarial est l'une des approches utilisées pour limiter ce problème. Un classifieur apprend à prédire la cible, tandis qu'un adversaire tente de retrouver les attributs sensibles à partir des représentations apprises par le modèle. En rendant cette récupération difficile, le classifieur apprend à moins s'appuyer sur ces attributs. En pratique, la méthode reste difficile à contrôler. La force de la contrainte d'équité est fixée à la main avant l'entraînement, sans mécanisme pour viser un niveau d'équité donné, et les attributs sensibles doivent être déclarés manuellement pour chaque jeu de données. Cet exposé présente une approche agentique du débiaisage adversarial. Le système identifie les attributs sensibles, mesure l'équité à chaque étape de l'entraînement et ajuste automatiquement la contrainte appliquée au modèle, en protégeant plusieurs attributs sensibles au cours d'une même exécution. L'équité n'est plus un réglage figé, mais une décision révisée tout au long de l'apprentissage. L'approche est évaluée sur le benchmark FFB et sur HIMS-Tunisia, un jeu de données issu de l'enquête nationale tunisienne sur la migration, puis comparée aux méthodes de référence.

Biographie courte Nahla Ben Amor

Nahla Ben Amor est professeure en informatique de gestion à l'Institut Supérieur de Gestion de Tunis (Université de Tunis). Elle a dirigé le Laboratoire de Recherche Opérationnelle, de Décision et de Contrôle de Processus (LARODEC) de 2013 à 2025. Elle est membre associée de l'Institut de Recherche en Informatique de Toulouse (IRIT) et membre du conseil scientifique de la Maison des Sciences de l'Homme et de la Société de Toulouse (MSHS-T). Elle est également auditrice principale senior certifiée ISO/IEC 42001 (Systèmes de management de l'intelligence artificielle).

Ses travaux de recherche portent sur les modèles graphiques, le raisonnement et la décision sous incertitude, l'apprentissage automatique et, plus récemment, l'IA agentique et l'IA responsable, incluant les questions d'équité algorithmique et de gouvernance des systèmes intelligents. Elle a encadré une vingtaine de thèses de doctorat et publié plus de 100 articles dans des revues et conférences internationales. Elle participe aux comités de programme de conférences internationales majeures en intelligence artificielle, notamment l'International Joint Conference on Artificial Intelligence (IJCAI), Uncertainty in Artificial Intelligence (UAI) et l'European Conference on Artificial Intelligence (ECAI). Elle intervient également comme relectrice pour des revues internationales telles que Fuzzy Sets and Systems (FSS) et International Journal of Approximate Reasoning (IJAR).

Adversarial Debiasing via Agentic Orchestration for Automated Fairness Enforcement

Machine learning models are increasingly used in sensitive areas such as credit scoring, recruitment, justice, and public services. They learn from historical data that carries the trace of past

*Page personnelle.

decisions, and can therefore reproduce the biases present in that data. Adversarial debiasing is one of the approaches used to limit this problem. A classifier learns to predict the target, while an adversary tries to recover the sensitive attributes from the representations learned by the model. By making this recovery more difficult, the classifier learns to rely less on those attributes. In practice, the method remains hard to control. The strength of the fairness constraint is set by hand before training, with no mechanism to target a given fairness level, and the sensitive attributes must be declared manually for each dataset. This talk presents an agentic approach to adversarial debiasing. The system identifies the sensitive attributes, measures fairness at each step of training, and automatically adjusts the constraint applied to the model, protecting several sensitive attributes within a single run. Fairness is no longer a fixed setting, but a decision revised throughout learning. The approach is evaluated on the FFB benchmark and on HIMS-Tunisia, a dataset from Tunisia's national migration survey, and compared with reference methods.

Short Bio Nahla Ben Amor

Nahla Ben Amor is a full Professor of Computer Science at the Institut Supérieur de Gestion de Tunis (University of Tunis). She served as Director of the Operations Research, Decision and Process Control Laboratory (LARODEC) from 2013 to 2025. She is an Associate Researcher at the Institut de Recherche en Informatique de Toulouse (IRIT) and a member of the Scientific Council of the Maison des Sciences de l'Homme et de la Société de Toulouse (MSHS-T). She is also a certified ISO/IEC 42001 Senior Lead Auditor (Artificial Intelligence Management Systems).

Her research interests include graphical models, reasoning and decision-making under uncertainty, machine learning and, more recently, agentic AI and responsible AI, including issues related to algorithmic fairness and the governance of intelligent systems. She has supervised around twenty PhD theses and published over 100 papers in international journals and conferences. She serves on the program committees of major AI conferences, including the International Joint Conference on Artificial Intelligence (IJCAI), Uncertainty in Artificial Intelligence (UAI) and the European Conference on Artificial Intelligence (ECAI). She also acts as a reviewer for leading journals such as Fuzzy Sets and Systems (FSS) and the International Journal of Approximate Reasoning (IJAR).

SESSION 1 : Incertitude et classification présidée par Henri Prade

Et si les classifieurs profonds oubliaient l'inconnu ? Auto-encodeurs et détection hors distribution

François Thievon[1], Arthur Hoarau[1,2]

[1] CentraleSupélec, Université Paris-Saclay, Metz, France
[2] Loria, CNRS, Université de Lorraine, Nancy, France

Résumé

La détection d'observations hors distribution (OOD) constitue un enjeu central pour le déploiement fiable des modèles d'apprentissage automatique, en particulier dans des systèmes critiques. En apprentissage profond, cette tâche est fréquemment abordée via la quantification d'incertitude épistémique, supposée élevée lorsque le modèle rencontre une entrée supposément éloignée de ses données d'entraînement. Cependant, des résultats récents montrent que de nombreuses méthodes échouent à séparer correctement incertitudes épistémique et aléatoire, ce qui dégrade les performances OOD.
Dans cet article, nous soutenons que cette difficulté provient en partie d'un biais inhérent à la classification : en optimisant la performance discriminative, un classifieur profond peut apprendre une représentation latente qui élimine précisément l'information de densité nécessaire à la détection OOD. Nous illustrons ce phénomène dans un cadre simulé, puis sur une expérience de classification d'images. Nos résultats montrent que, pour une même méthode de quantification d'incertitude, la détection OOD est significativement améliorée lorsque l'on travaille dans l'espace latent d'un auto-encodeur plutôt que dans celui d'un classifieur.

Mots-clés

Quantification d'Incertitude, Détection hors distribution, Classification.

Abstract

Out-of-distribution (OOD) detection is a central challenge for the reliable deployment of machine learning models, especially in safety-critical systems. In deep learning, this task is often addressed through epistemic uncertainty quantification, which is assumed to be high when the model encounters an input that is supposedly far from its training data. However, recent results show that many methods fail to properly disentangle epistemic and aleatoric uncertainty, which in turn degrades OOD detection performance.
In this paper, we argue that this difficulty partly stems from a bias inherent to classification itself : by optimizing discriminative performance, a deep classifier may learn a latent representation that precisely removes the density information required for OOD detection. We illustrate this phenomenon in a controlled simulated setting, and then on an image classification experiment. Our results show that, for the same uncertainty quantification method, OOD detection is significantly improved when operating in the latent space of an autoencoder rather than that of a classifier.

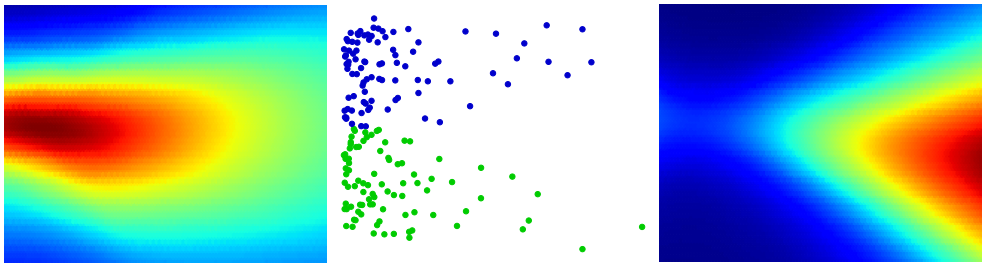
Keywords

Uncertainty Quantification, Classification, OOD Detection.

1 Introduction

La recherche en apprentissage automatique s'est longtemps attachée à améliorer la performance prédictive des modèles. Cependant, avec les avancées récentes en intelligence artificielle (vision, recommandation, modèles de langage, etc.) et son intégration dans un nombre croissant de systèmes parfois critiques, de nouveaux défis émergent. Une attention grandissante est désormais portée aux coûts écologiques, économiques et sociétaux de l'IA [21], ainsi qu'à la demande de modèles plus frugaux. Pour des modèles déjà capables d'atteindre des performances proches de l'optimal, il devient crucial de capturer et de rapporter l'incertitude associée aux prédictions locales les plus sensibles, même si elles sont rares. Une seule défaillance peut être fatale dans des systèmes critiques pour la sécurité, tels que la conduite autonome, les applications médicales ou spatiales. La récente révolution des grands modèles de langage (LLM[1]) et leur démocratisation ont également soulevé des inquiétudes majeures concernant la transparence des prédictions, notamment face à l'excès de confiance que certains modèles affichent dans leurs prédictions. Dans la plupart des applications, des modèles dits "boîte noire" sont déployés, et la recherche vise en partie à rendre plus transparentes les prédictions de ces modèles parfois très complexes.
Dans ce contexte, l'évaluation des modèles d'apprentissage automatique doit aller au-delà de la performance prédictive brute et prendre explicitement en compte l'incertitude et la confiance associées aux prédictions. La détection d'observations hors distribution [17, 13] (détection OOD[2] en phase de test) vise précisément à identifier les entrées pour lesquelles le modèle n'a pas été correctement entraîné [6, 29]. Par exemple, un modèle assistant un médecin et entraîné à catégoriser la pathologie de nouveaux patients devrait être

1. LLM pour *Large Language Model*.
2. OOD pour *Out-of-distribution.*



(a) Incertitude Aléatoire (b) Jeu de données (c) Incertitude Épistémique

FIGURE 1 – Décomposition d'incertitude par apprentissage profond bayésien et approximation de Laplace. L'incertitude aléatoire est présentée en figure 1a, et l'incertitude épistémique en figure 1c, avec un dégradé allant du rouge au bleu, correspondant respectivement à une incertitude élevée et faible.

capable de signaler son incertitude lorsqu'un patient présentant une maladie inconnue (comme le COVID-19 lors de son émergence) se présente. De même, dans un contexte de conduite autonome, la reconnaissance de panneaux de signalisation peut être fortement dégradée par la présence de neige ou de brouillard, absents du jeu d'entraînement.

La quantification d'incertitude est un champ d'étude de l'apprentissage automatique qui s'est en partie attelé à adresser cette question de détection OOD. Dans la littérature, on distingue généralement deux grands types d'incertitude [8] : l'incertitude aléatoire (liée aux données) et l'incertitude épistémique (liée au modèle). L'incertitude aléatoire provient de la nature stochastique du processus générateur des données, tandis que l'incertitude épistémique est associée à un manque de connaissance. La première est généralement considérée comme irréductible, alors que la seconde peut être atténuée par l'acquisition de données supplémentaires. Ces deux types d'incertitude sont illustrés en figure 1, à l'aide d'un modèle d'apprentissage bayésien profond et d'une approximation de Laplace [12], sur un jeu de données synthétique. L'incertitude aléatoire permet d'identifier les zones où la prédiction est intrinsèquement difficile : dans ces régions, les deux classes du jeu de données se chevauchent effectivement. À l'inverse, une incertitude épistémique élevée traduit un manque de connaissance du modèle. Elle correspond à des régions de l'espace des entrées insuffisamment couvertes par les données d'entraînement, pour lesquelles il serait donc risqué de se fier à la prédiction du modèle. C'est précisément ce second type d'incertitude qui est généralement mobilisé en quantification d'incertitude pour détecter des observations hors distribution.

Afin de dissocier l'incertitude aléatoire de l'incertitude épistémique, de nombreuses méthodes ont été proposées ces dernières années [9]. Cependant, cet article de position se veut le plus ouvert possible ; c'est pourquoi nous avons choisi de présenter ces méthodes dans une section dédiée, que nous considérons indépendante du reste de l'article.

L'utilisation de l'incertitude épistémique pour la détection hors distribution (OOD) est largement admise et s'est montrée particulièrement efficace. Cependant, un nombre croissant de travaux se montrent critiques vis-à-vis de cette approche [16], en particulier dans le cadre de l'apprentissage profond. Des travaux récents ont notamment mis en évidence une corrélation (trop) importante entre les composantes aléatoire et épistémique [19] pour la plupart des méthodes considérées.

Dans cet article, nous soutenons qu'il existe un biais inhérent, non pas lié à la méthode de quantification d'incertitude employée, mais à la tâche de classification elle-même, et que ce biais peut en partie expliquer les échecs récents de la détection OOD en apprentissage profond. En conséquence, la comparaison des méthodes de quantification d'incertitude épistémique peut s'avérer secondaire, dès lors que l'information déterminante pour la détection OOD est éliminée en amont par la tâche de classification elle-même. Nous montrons également que ce phénomène peut être partiellement atténué par l'utilisation d'auto-encodeurs, qui préservent mieux la densité dans l'espace latent et permettent ainsi une détection OOD plus fiable. Cet avantage s'accompagne toutefois d'un coût en termes de performances : les représentations latentes issues d'un auto-encodeur sont généralement moins discriminantes que les sorties d'un classifieur. Dans cet article, nous utilisons des modèles simples et n'employons pas, par exemple, d'auto-encodeurs variationnels [4]. Bien que leurs avantages soient nombreux, ils sortent du cadre de ce travail. Ici, c'est la différence de tâche entre classifieur et auto-encodeur qui constitue le point central. En résumé, nous souhaitons attirer l'attention du lecteur sur le fait que la recherche de performances maximales, en particulier en classification, induit un coût rarement explicité : la réduction de la capacité du modèle à détecter les données hors distribution. Autrement dit, les modèles les plus performants ont tendance à "oublier leur ignorance".

L'article est organisé comme suit ; la section 2 introduit les notions de quantification d'incertitude appliquées à la détection OOD. Le reste du papier est volontairement plus accessible et peut être lu indépendamment de cette section. La section 3 présente une mise en évidence empirique du problème, souligné dans cet article, de détection OOD en classification profonde. La section 4 propose une évaluation expérimentale sur un problème de classification d'images. Enfin, la section 5 conclut cet article.

2 Détection hors distribution & Quantification d'incertitude

Dans cette section, nous introduisons la détection hors distribution (OOD) à travers la notion d'incertitude épistémique présentée en introduction. La section 2.1 présente les principales méthodes permettant de capturer et de représenter les incertitudes d'un modèle de classification, tandis que la section 2.2 explicite la quantification d'incertitude épistémique dans les prédictions d'un modèle d'apprentissage automatique.

2.1 Quantification d'incertitude

Dans la littérature, on distingue généralement deux types d'incertitude [8] : l'incertitude aléatoire (liée aux données) et l'incertitude épistémique (liée au modèle). L'incertitude aléatoire provient du caractère stochastique du processus générateur des données, tandis que l'incertitude épistémique est associée à un manque de connaissance. La première est généralement considérée comme irréductible, alors que la seconde peut être atténuée par l'acquisition de données supplémentaires. Afin de dissocier ces deux composantes, de nombreuses méthodes ont été proposées ces dernières années [9]. Ces approches peuvent être regroupées en quatre grandes familles.

Les méthodes bayésiennes calculent une distribution a posteriori sur les paramètres d'un modèle et construisent une distribution de second ordre sur les probabilités de classe via la moyenne bayésienne de modèles (*Bayesian model averaging*) [12, 3, 27] :

$$p(h \mid \mathcal{D}) \propto p(h)p(\mathcal{D} \mid h). \tag{1}$$

La quantification d'incertitude en apprentissage automatique repose traditionnellement sur l'apprentissage bayésien, où $p(h \mid \mathcal{D})$ désigne la densité de probabilité a posteriori d'un modèle $h \in \mathcal{H}$ conditionnellement à un jeu de données $\mathcal{D}$. Intuitivement, la dispersion de $p(h \mid \mathcal{D})$ reflète l'état de connaissance du modèle, et donc son incertitude épistémique : plus cette distribution est concentrée (c'est-à-dire qu'un modèle est nettement plus probable que les autres), moins l'apprenant est épistémiquement incertain quant au fait que h^* soit le modèle optimal. Il est crucial de noter que le fait que h^* soit le meilleur modèle n'implique pas que sa prédiction $\hat{y} = h^*(x)$, ou encore $p_{h^*}(y \mid x)$, soit certaine du point de vue aléatoire. Autrement dit, un modèle peut être épistémiquement confiant tout en reconnaissant que la tâche est intrinsèquement aléatoirement difficile. Malgré de solides fondements théoriques, l'apprentissage bayésien a également fait l'objet de critiques substantielles, notamment en raison de ses hypothèses parfois restrictives et de la nécessité de recourir à des approximations empiriques afin de rendre l'inférence *tractable*.

Les méthodes d'ensembles capturent l'incertitude épistémique à travers la diversité des estimateurs. Cette diversité peut être obtenue en *randomisant* les données d'entraînement par *bootstrapping*, en variant l'architecture ou les mécanismes de régularisation (par exemple via *MC Dropout* ou *DropConnect* [18]), ou encore en *randomisant* l'optimisation, comme pour *Deep Ensembles* [14]. L'incertitude épistémique est alors élevée lorsque le désaccord entre estimateurs est important, tandis que l'incertitude aléatoire est élevée lorsque l'entropie moyenne des estimateurs est importante. Il a été montré que certaines de ces techniques peuvent être interprétées comme des approximations de la distribution a posteriori de second ordre mentionnée précédemment [12]. Bien qu'intuitive et largement applicable, cette famille de méthodes a également été critiquée quant à sa justification théorique.

L'*Evidential Deep Learning* (EDL) adopte une perspective fréquentiste et vise à minimiser à la fois l'erreur du modèle et l'incertitude épistémique, cette dernière étant représentée par une probabilité de second ordre, sans nécessiter la spécification d'un a priori ou d'un a posteriori sur les paramètres. Dans le cas de la classification multi-classes, la distribution de Dirichlet est généralement utilisée [23, 26]. Malgré des performances empiriques raisonnables, cette famille récente de méthodes a fait l'objet de nombreuses critiques [1]. En particulier, certains comportements attendus se sont révélés impossibles à satisfaire [11], conduisant à la conclusion générale que l'EDL n'est pas adapté à une quantification correcte de l'incertitude épistémique.

Les méthodes fondées sur la densité/distance modélisent typiquement les incertitudes aléatoire et épistémique directement dans l'espace des variables [20, 7]. Ces approches peuvent également construire une distribution de second ordre dans une étape de post-traitement [2]. En pratique, l'incertitude épistémique associée à une prédiction est élevée lorsque sa vraisemblance est faible pour toutes les classes du jeu de données, tandis que l'incertitude aléatoire est élevée lorsque l'instance est susceptible d'appartenir à plusieurs classes. Le principal inconvénient de ces méthodes est que l'incertitude épistémique peut être non bornée ou insuffisamment justifiée théoriquement, bien qu'elles puissent au moins produire un classement cohérent des prédictions.

En raison des limites intrinsèques de la théorie des probabilités classique pour représenter certaines formes d'ignorance et d'imprécision, plusieurs cadres mathématiques alternatifs ont été développés [5]. La théorie des fonctions de croyance, par exemple, permet de représenter différents niveaux d'ignorance. Les probabilités imprécises permettent la construction d'ensembles crédaux, formalisant des bornes inférieures et supérieures pour chaque probabilité. De même, la théorie des sous-ensembles flous fournit un formalisme permettant de représenter des frontières de classes graduelles.

2.2 Incertitude épistémique

Soit (X, Y) une paire de variables aléatoires[3], où X prend ses valeurs dans un espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^r$ et Y dans l'espace des sorties $\mathcal{Y} = \{1, \ldots, K\}$. Les données sont supposées générées selon une distribution jointe inconnue P sur $\mathcal{X} \times \mathcal{Y}$. Un jeu de données d'entraînement est donné par

3. Nous omettons ici la notion de vecteur aléatoire par souci de simplicité.

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, constitué de n réalisations i.i.d. selon P. Un classifieur est une fonction $h : \mathcal{X} \to \mathcal{Y}$ appartenant à une classe d'hypothèses $\mathcal{H}$ et produisant une distribution conditionnelle $p_h(y \mid x)$.

L'outil optimal [10] pour la détection hors distribution d'une nouvelle instance x est la densité $p(x)$, ou, en classification, la densité conditionnelle $p(x \mid y)$. Une instance présentant une densité conditionnelle très faible pour toutes les classes a peu de chances d'appartenir à la distribution d'entraînement. Cependant, le processus générateur étant inconnu, cette approche est irréalisable en pratique. Pour pallier cette difficulté, il est possible d'utiliser l'incertitude épistémique associée à la prédiction $p_h(y \mid x)$: une incertitude épistémique élevée indique que le modèle a affaire à une entrée pour laquelle il n'a pas été spécifiquement entraîné.

Considérons un prédicteur de second ordre, défini comme un ensemble de modèles $\Delta \subseteq \mathcal{H}$. Dans la pratique, Δ peut correspondre à un ensemble d'estimateurs (méthodes d'ensembles), à un ensemble convexe de probabilités (ensembles crédaux), ou à un espace paramétrique (cadre bayésien). Dans la section 4, Δ correspond à une forêt aléatoire d'arbres de décision.

Pour chaque entrée $x \in \mathcal{X}$, chaque hypothèse $h \in \Delta$ produit une probabilité prédictive $p_h(y \mid x)$, agrégée sous la forme de la prédiction du modèle de second ordre $p_\Delta(y \mid x)$. Une décomposition entropique répandue de l'incertitude s'écrit :

$$\underbrace{H[p_\Delta(y \mid x, \mathcal{D})]}_{\text{Inc. Totale}} = \underbrace{\mathbb{E}[H[p_h(y \mid x)]]}_{\text{Inc. Aléatoire}} + \underbrace{I(y, h \mid x, \mathcal{D})}_{\text{Inc. Epistémique}}, \tag{2}$$

où $H(\cdot)$ et $I(\cdot \mid \cdot)$ désignent respectivement l'entropie de Shannon [25] et l'information mutuelle.

Dans le cas discret des forêts aléatoires [24], l'incertitude aléatoire IA_{ent} est la moyenne des entropies prédictives des M estimateurs h_m :

$$\text{IA}_{ent}(x) = -\frac{1}{M} \sum_{m=1}^{M} \sum_{y \in \mathcal{Y}} p_{h_m}(y \mid x) \log \left(p_{h_m}(y \mid x) \right). \tag{3}$$

L'incertitude totale IT_{ent} correspond à l'entropie de la prédiction moyenne :

$$\text{IT}_{ent}(x) = - \sum_{y \in \mathcal{Y}} \left(\frac{1}{M} \sum_{m=1}^{M} p_{h_m}(y \mid x) \right) \log \left(\frac{1}{M} \sum_{m=1}^{M} p_{h_m}(y \mid x) \right). \tag{4}$$

L'incertitude épistémique IE_{ent} est alors obtenue par différence entre l'incertitude totale et l'incertitude aléatoire :

$$\text{IE}_{ent}(x) = \text{IT}_{ent}(x) - \text{IA}_{ent}(x), \tag{5}$$

et représente l'incertitude liée au désaccord entre les arbres, au sein de la forêt.

Afin de diversifier les mesures, nous considérons également une seconde définition de l'incertitude épistémique fondée sur la variance de la probabilité prédictive [22] :

$$\text{IE}_{var}(x) = \sum_{y \in \mathcal{Y}} \text{Var}\left[p_h(y \mid x) \right], \tag{6}$$

qui, dans le cas des forêts aléatoires, s'écrit :

$$\text{IE}_{var}(x) = \frac{1}{M} \sum_{y \in \mathcal{Y}} \sum_{m=1}^{M} \left(p_{h_m}(y \mid x) - \left(\frac{1}{M} \sum_{i=1}^{M} p_{h_i}(y \mid x) \right) \right)^2, \tag{7}$$

et qui représente la variance entre les prédictions des arbres, au sein de la forêt.

Dans la section 4, nous utiliserons donc IE_{ent} et IE_{var} dans un objectif de détection OOD. La section suivante, quant à elle, ne fait pas encore appel à ces mesures et se concentre sur la mise en évidence du phénomène dans un cadre simulé.

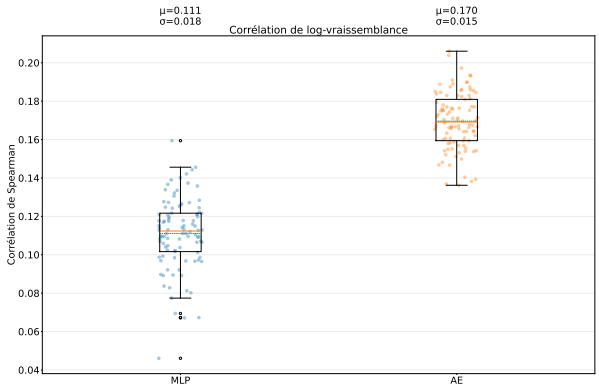


FIGURE 2 – Corrélations de Spearman entre la densité dans l'espace d'entrée et la densité dans l'espace latent du MLP et de l'AE.

3 Quand la classification atteint ses limites

Cette section met en évidence le phénomène central de cet article : en classification profonde, l'information pertinente pour détecter les données hors distribution peut être progressivement éliminée par l'apprentissage, au profit de représentations uniquement optimisées pour la séparation des classes.

En apprentissage automatique classique, de nombreuses approches de détection OOD par quantification d'incertitude (souvent plus robustes que leurs équivalents profonds) opèrent directement dans l'espace d'entrée. Une instance est alors considérée comme hors distribution au sens où elle est atypique dans l'espace des variables. Toutefois, lorsque les données sont complexes ou de grande dimension (images, texte, etc.), travailler directement dans l'espace d'entrée devient rapidement impossible. Les réseaux

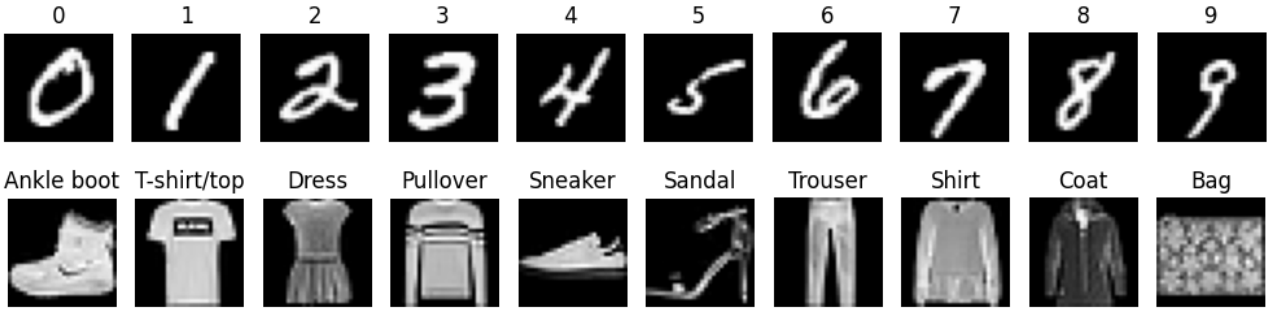


FIGURE 3 – Jeu de données en distribution MNIST (haut) et jeu de données hors distribution Fashion-MNIST (bas).

de neurones profonds[4] jouent alors le rôle d'extracteurs de caractéristiques : couche après couche, ils transforment la donnée afin de produire une représentation de plus en plus abstraite, sur laquelle la prédiction finale devient simple.
Pour illustrer le problème, considérons une classification binaire d'images entre voitures et bicyclettes. De manière schématique, un espace latent intermédiaire peut capturer un concept simple tel que le nombre de roues. La séparation devient alors presque linéaire : moins de deux roues implique une bicyclette, sinon une voiture. Dans ce contexte, une motocyclette, pourtant clairement hors distribution dans l'espace d'entrée, peut devenir *en distribution* dans l'espace latent associé au nombre de roues. Le modèle peut alors prédire avec une confiance élevée qu'il s'agit d'une bicyclette[5].
Ce mécanisme illustre un point clé : en classification, le modèle apprend principalement les invariances et les abstractions nécessaires à maximiser la performance sur la tâche. En conséquence, la représentation peut perdre des informations pourtant cruciales pour détecter des entrées atypiques. La section suivante propose une illustration contrôlée de ce phénomène.

3.1 Illustration simulée

Afin de contrôler explicitement la densité $p(x)$ dans l'espace d'entrée, nous générons les données selon un mélange de gaussiennes :

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x \mid \mu_k, \Sigma_k), \tag{8}$$

où $\mathcal{N}(x \mid \mu_k, \Sigma_k)$ est la densité d'une gaussienne multivariée et π_k est le *prior* associé à la classe k. Dans nos expériences, la distribution est uniforme sur les classes, i.e. $\pi_k = \frac{1}{K}$. À chaque itération, les moyennes μ_k et les matrices de covariance Σ_k sont tirées aléatoirement. Nous utilisons des matrices de covariance de rang faible et faiblement bruitées[6].

Pour chaque itération, nous générons 5000 données d'entraînement et 5000 données de test, selon 100 variables et 3 classes. Un perceptron multi-couches (MLP) et un auto-encodeur (AE) sont entraînés respectivement à prédire la classe et à reconstruire la donnée d'entrée. Les deux modèles disposent d'un espace latent de dimension 10.

Le jeu de test est ensuite projeté dans l'espace latent du MLP et de l'AE. Nous comparons la log-vraisemblance négative $-\log p(x)$ dans l'espace d'entrée à une estimation de $-\log p(z)$ dans l'espace latent, où z désigne la représentation latente de x. Pour estimer $\log p(z)$, nous utilisons une estimation de densité par noyau gaussien (KDE), avec une largeur de bande fixée à $0.2\sqrt{10}$.

La figure 2 présente les résultats obtenus sur 100 itérations. La corrélation de Spearman entre la densité dans l'espace d'entrée et celle dans l'espace latent est reportée pour le MLP[7] (gauche) et pour l'AE (droite). Lorsque qu'une instance tend à être hors distribution dans l'espace d'entrée, elle est nettement moins souvent associée à une faible densité dans l'espace latent du MLP que dans celui de l'AE.

C'est précisément le phénomène que nous souhaitons souligner : même si un AE classique ne garantit pas la conservation de densité (cf. section 5 pour des alternatives), celle-ci est empiriquement mieux préservée que dans le cas d'un classifieur profond. La section suivante étudie ce phénomène sur un cas concret de détection OOD en classification d'images.

4 Expériences

Dans cette section, nous montrons que l'utilisation de classifieur profond dégrade les performances de détection OOD sur le jeu de données MNIST [15].

Le modèle est entraîné sur MNIST, puis des données hors distribution issues de Fashion-MNIST [28] sont injectées lors de la phase de test (cf. figure 3).

Deux architectures sont étudiées : un perceptron multi-couches (MLP) et un réseau convolutif (CNN). Nous considérons ainsi quatre modèles : deux classifieurs (MLP et CNN) et deux auto-encodeurs (AE-MLP et AE-CNN).

4. Nous parlons ici des réseaux de neurones que nous utilisons dans l'article, à savoir des réseaux convolutifs et des perceptrons multi-couches.
5. Dans cet exemple, un avion à 3 ou 6 roues, OOD dans l'espace d'entrée, resterait OOD dans l'espace de représentation latent. Toute l'information de détection OOD n'est pas perdue dans l'extraction de caractéristiques.
6. Matrices de rang 5.
7. À titre indicatif, l'exactitude moyenne du MLP est de 0.94.

4.1 Protocole expérimental

Le MLP suit l'architecture $784 \to 512 \to 256 \to 128 \to 64 \to 32 \to 10$. Toutes les activations sont des ReLU, à l'exception de la dernière couche, linéaire. Le modèle est entraîné avec une perte d'entropie croisée, un pas d'apprentissage de 10^{-3}, une taille de batch de 128, pendant 4 epochs. Le CNN suit l'architecture $\text{Conv}(32, 3 \times 3) \to \text{Conv}(64, 3 \times 3) \to \text{MaxPool}(2) \to \text{Conv}(128, 3 \times 3) \to \text{MaxPool}(2) \to 128 \times 7 \times 7 \to 256 \to 128 \to 32 \to 10$. Les auto-encodeurs AE-MLP et AE-CNN utilisent le même encodeur que leurs homologues classifieurs, et un décodeur symétrique après le goulot d'étranglement en dimension 32. Ils sont entraînés avec une perte quadratique moyenne, pendant 10 epochs[8].

Les quatre modèles sont entraînés sur les 60 000 images du jeu d'entraînement de MNIST (28×28, normalisées). Pour chacun d'eux, une représentation latente de dimension 32 est extraite. À partir de ces espaces latents (MLP, CNN, AE-MLP et AE-CNN), une forêt aléatoire composée de 200 arbres est entraînée.

Les données MNIST de test (10 000 échantillons en distribution) ainsi que les données Fashion-MNIST de test (10 000 échantillons hors distribution) sont ensuite projetées dans l'espace latent de chaque modèle. La détection hors distribution (OOD) est réalisée dans cet espace à l'aide de la forêt aléatoire, en s'appuyant sur une mesure d'incertitude épistémique calculée selon les équations (5) et (7). Une valeur élevée de l'incertitude épistémique (IE_{ent} ou IE_{var}) est supposée indiquer une instance hors distribution.

4.2 Comparaison de performance

À titre indicatif, le MLP atteint une exactitude de 0.975 et le CNN une exactitude de 0.988 sur le jeu de test MNIST. Les résultats qualitatifs obtenus en sortie des auto-encodeurs AE-MLP et AE-CNN sont présentés en figure 4.

Les figures 5 et 6 présentent respectivement les incertitudes épistémiques calculées par décomposition entropique (5) et par variance (7). Pour chaque modèle, l'histogramme du nombre d'instances de test est représenté en fonction de l'incertitude épistémique, en distinguant les données MNIST (en distribution) et Fashion-MNIST (hors distribution).

Pour les deux classifieurs (MLP et CNN), les distributions d'incertitude associées à MNIST et Fashion-MNIST se chevauchent partiellement très tôt (des instances OOD sont considérées en distribution). Autrement dit, les représentations latentes apprises pour la classification ne permettent pas toujours de distinguer clairement les instances hors distribution. À l'inverse, lorsque la détection est réalisée dans l'espace latent des auto-encodeurs, la séparation entre les deux ensembles est nettement plus marquée, ce qui rend possible un seuillage plus robuste.

Il est toutefois important de noter que, pour les auto-encodeurs, les données OOD et les données en distribution présentent un chevauchement systématique. Par conséquent, il est impossible de distinguer MNIST de Fashion-MNIST sur la seule base de la prédiction du modèle. Néanmoins, nous soutenons que ce phénomène ne traduit pas une défaillance de la détection OOD. Il reflète plutôt la capacité du modèle à identifier, au sein même du jeu de données de référence, des instances qui présentent des caractéristiques OOD. La Figure 7 illustre ce point en présentant les trois instances de test de MNIST pour lesquelles l'AE-MLP manifeste la plus forte incertitude épistémique. Bien qu'appartenant formellement au jeu de données MNIST, ces exemples présentent des caractéristiques clairement OOD.

Ces observations sont cohérentes avec l'hypothèse défendue dans cet article. Les classifieurs apprennent une représentation latente fortement contrainte par l'objectif discriminatif, qui favorise la séparation des classes MNIST mais ne préserve pas nécessairement la structure de densité de l'espace d'entrée. Ainsi, certaines instances de Fashion-MNIST peuvent être projetées dans des régions latentes similaires à celles de MNIST, rendant la détection OOD difficile, même lorsque l'on dispose d'une mesure d'incertitude épistémique raisonnable.

À l'inverse, les auto-encodeurs sont entraînés à reconstruire l'image d'entrée, ce qui les contraint à encoder un spectre plus large d'informations, y compris des facteurs de variation peu utiles pour la classification des chiffres (texture, contours, style d'écriture, structure locale, etc.). Cette préservation plus riche des caractéristiques visuelles semble favoriser une meilleure séparation entre données en distribution et hors distribution dans l'espace latent, et explique la nette amélioration observée en détection OOD.

Enfin, le choix de la mesure d'incertitude épistémique, décomposition entropique (5) ou variance (7), influence peu qualitativement les résultats, et la conclusion reste inchangée.

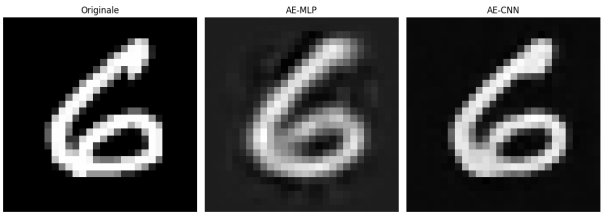


FIGURE 4 – Résultats obtenus en sortie des auto-encodeurs. Chaque modèle encode puis décode l'image d'entrée pour la reconstruire.

5 Discussion & Conclusion

Dans cet article, nous avons montré que la détection d'instances hors distribution par quantification d'incertitude peut s'avérer significativement plus difficile, voire pratiquement impossible, lorsqu'elle est réalisée dans des espaces latents issus de classifieurs profonds, comparativement à des espaces latents issus d'auto-encodeurs. Ce phénomène s'explique par la capacité des classifieurs à apprendre des représentations abstraites optimisées pour la séparation des classes, sans contrainte explicite de conservation de la

8. Le nombre d'epochs est choisi en fonction de la convergence observée.

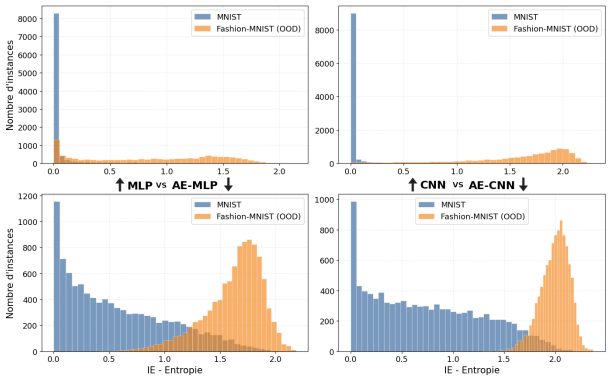


FIGURE 5 – Histogramme de l'incertitude épistémique (IE) selon une décomposition entropique. Les classifieurs sont présentés sur la ligne supérieure et les auto-encodeurs sur la ligne inférieure.

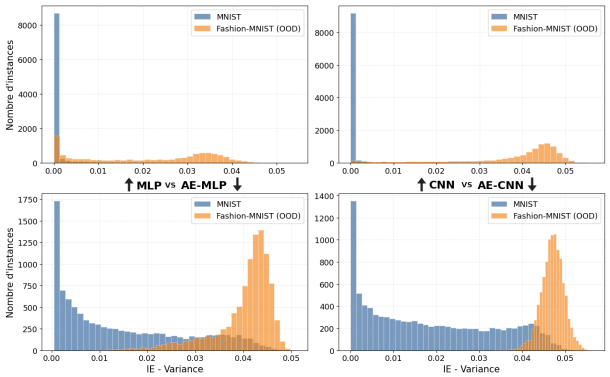


FIGURE 6 – Histogramme de l'incertitude épistémique (IE) selon une décomposition par variance. Les classifieurs sont présentés sur la ligne supérieure et les auto-encodeurs sur la ligne inférieure.

structure de densité dans l'espace d'entrée.

Certaines approches existent pour préserver davantage cette information, par exemple en ajoutant des contraintes sur les distances dans l'espace latent, en utilisant des *normalizing flows*, ou en imposant des contraintes de type *bi-Lipschitz*. Dans tous les cas, ces solutions reviennent à introduire un compromis entre performance de classification et capacité à détecter les données hors distribution, compromis illustré dans nos expériences par la comparaison entre classifieurs et auto-encodeurs.

Nous ne proposons pas de preuve formelle, car nous pensons qu'il n'existe pas de garantie théorique générale : un classifieur peut, dans certains cas, apprendre une représentation latente qui préserve mieux la densité d'origine qu'un auto-encodeur. L'objectif n'est donc pas d'affirmer que les auto-encodeurs sont systématiquement supérieurs, mais de souligner que la détection OOD est intrinsèquement liée au type de représentation appris, et donc à la tâche d'apprentissage.

Pour conclure, nous soutenons qu'évaluer des méthodes de

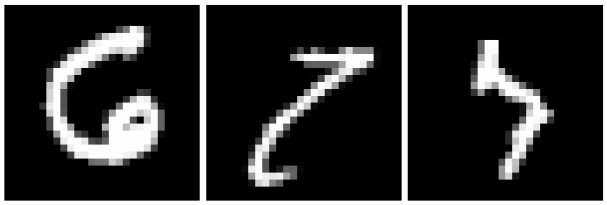


FIGURE 7 – Instances de test de MNIST pour lesquelles l'AE-MLP présente la plus forte incertitude épistémique. Ces instances sont déjà OOD dans le jeu de données témoin.

quantification d'incertitude épistémique en apprentissage profond uniquement à travers des benchmarks de détection hors distribution constitue une approche incomplète. Certaines idées peuvent être pertinentes au niveau de la mesure d'incertitude, mais les performances peuvent être fortement limitées si l'espace de représentation latent a déjà éliminé l'information nécessaire à la détection OOD.

Plus généralement, ce travail vise à attirer l'attention sur un coût rarement explicité de l'optimisation de la performance en classification. Gagner quelques points d'exactitude (*accuracy*) peut se faire au prix d'une représentation plus compressée et plus abstraite, dans laquelle l'information nécessaire à la détection des cas atypiques est partiellement supprimée. Autrement dit, l'amélioration marginale des performances peut réduire la capacité du modèle à reconnaître ses propres limites, et donc à signaler de manière fiable *quand il ne sait pas*.

Remerciements

Nous souhaitons remercier Gauthier Justo et Théo Beaumont, étudiants à CentraleSupélec, pour leur implication dans le projet de fin d'études sur lequel s'appuient les travaux complémentaires présentés dans cet article. Nous remercions également Vincent Lemaire d'Orange Research pour sa participation à la relecture et à la préparation de la version camera-ready de cet article.

Références

[1] Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2078–2091. PMLR, 23–29 Jul 2023.

[2] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network : Uncertainty estimation without ood samples via density-based pseudo-counts. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1356–1367. Curran Associates, Inc., 2020.

[3] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1184–1193. PMLR, 10–15 Jul 2018.

[4] P. Kingma Diederik and Welling Max. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4) :307–392, 11 2019.

[5] Zhen Guo, Zelin Wan, Qisheng Zhang, Xujiang Zhao, Qi Zhang, Lance M. Kaplan, Audun Jøsang, Dong H. Jeong, Feng Chen, and Jin-Hee Cho. A survey on uncertainty reasoning and quantification in belief theory and its application to deep learning. *Information Fusion*, 101, 2024.

[6] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*, February 2017.

[7] Arthur Hoarau, Vincent Lemaire, Yolande Le Gall, Jean-Christophe Dubois, and Arnaud Martin. Evidential uncertainty sampling strategies for active learning. *Machine Learning*, 113(9) :6453–6474, June 2024.

[8] Stephen C. Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2), 1996. Treatment of Aleatory and Epistemic Uncertainty.

[9] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning : An introduction to concepts and methods. *Machine Learning*, 110 :457–506, 2021.

[10] Joonas, Joonas Rve, Karl Kaspar Haavel, and Meelis Kull. Probability Density from Latent Diffusion Models for Out-of-Distribution Detection. In *ECAI 2025*, pages 5027–5034. IOS Press, 2025.

[11] Mira Jürgens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods ? In *Proceedings of the 41st International Conference on Machine Learning*, International Conference on Machine Learning 24. JMLR.org, 2024.

[12] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision ? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[13] B. Lakshminarayanan. Practical tutorial on uncertainty and out-of-distribution robustness in deep learning, google research, link : https ://www.gatsby.ucl.ac.uk/b̃alaji/balaji-odsc-talk.pdf, 2024.

[14] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems*, pages 6402–6413, 2017.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11) :2278–2324, 1998.

[16] Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. *Position : Supervised Classifiers Answer the Wrong Questions for OOD Detection*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 13–19 Jul 2025.

[17] Shuo Lu, Yingsheng Wang, Lijun Sheng, Lingxiao He, Aihua Zheng, and Jian Liang. Out-of-distribution detection : A task-oriented survey of recent advances. *ACM Comput. Surv.*, 58(2), September 2025.

[18] Aryan Mobiny, Pengyu Yuan, Supratik Moulik, Naveen Garg, Carol wu, and Hien Nguyen. Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Scientific Reports*, 11, 03 2021.

[19] Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement : Specialized uncertainties for specialized tasks. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[20] Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111 :89–122, jan 2022.

[21] Billy Perrigo. Exclusive : The $2 Per Hour Workers Who Made ChatGPT Safer, January 2023.

[22] Y. Sale, P. Hofman, L. Wimmer, Hüllermeier E., and T. Nagler. Second-order uncertainty quantification : Variance-based measures., 2023.

[23] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[24] Mohammad Hossein Shaker and Eyke Hüllermeier. Aleatoric and epistemic uncertainty with random forests. In Michael R. Berthold, Ad Feelders, and Georg Krempl, editors, *Advances in Intelligent Data Analysis XVIII*, pages 444–456, Cham, 2020. Springer International Publishing.

[25] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3) :379–423, 1948.

[26] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks : A survey on evidential deep learning methods for uncertainty estimation. *Transactions of Machine Learning Research*, 2023.

[27] Amaan Valiuddin, Ruud Van Sloun, Christiaan Viviers, Peter H.N. de With, and Fons van der Sommen. A review of bayesian uncertainty quantification in deep probabilistic image segmentation. *Transactions on Machine Learning Research*, 2025.

[28] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST : a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv e-prints*, page arXiv :1708.07747, August 2017.

[29] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized Out-of-Distribution Detection : A Survey. *International Journal of Computer Vision*, 132(12) :5635–5662, December 2024.

Annotation linguistique sous incertitude structurelle : vers un modèle ensembliste basé sur les fonctions de croyance

Anna Pappa[1,2], Pierre Pomeret-Coquot[3], Francis Faux[2], Revekka Kyriakoglou[1]

[1] LIASD, Université Paris 8, 2 rue de la liberté, 93526 Saint Denis, France
[2] IRIT, Université de Toulouse, CNRS, Toulouse INP, Toulouse, France
[3] IRIT, Université de Toulouse, CNRS, Toulouse INP, UT Capitole, IUT de Rodez, Toulouse, France

ap@up8.edu – francis.faux@irit.fr – pierre.pomeret@irit.fr – kyriakoglou@up8.edu

Résumé

La majorité des travaux en annotation linguistique modélisent l'assignation d'une étiquette comme une application déterministe associant chaque instance à une unique catégorie. Cette hypothèse suppose que les catégories sont indépendantes les unes des autres et qu'une même unité ne peut appartenir qu'à une seule catégorie.

À partir d'un cas d'étude portant sur l'annotation fonctionnelle de prompts produits en contexte d'apprentissage, nous montrons que dans certains schémas catégoriels, l'indétermination est structurelle et provient du recouvrement partiel des catégories.

Nous étudions comment la théorie des fonctions de croyance permet de décrire avec exactitude les différents ensembles d'étiquettes proposés par les annotateurs, en représentant explicitement l'incertitude et les désaccords.

Cette annotation ensembliste, bien que non univoque, conserve l'information interprétative et permet d'extraire plusieurs sous-corpus annotés et d'en mesurer la qualité. Nous identifions également des regroupements pertinents de catégories d'étiquettes, permettant de réduire l'incertitude sans perdre en granularité sémantique.

Le corpus annoté obtenu est donc exact et indépendant des tâches d'apprentissage auxquelles il est destiné, tout en offrant un cadre susceptible d'améliorer la précision des modèles et d'affiner leur évaluation.

Mots-clés

Annotation TALN Incertitude Fonctions de croyance

Abstract

Most works in linguistic annotation models label assignment as a deterministic function that associates each instance to a single category. This assumption implies that categories are independent from one another and that a given annotation unit belongs to only one label.

Based on a case study involving the functional annotation of prompts produced in a learning context, we show that in certain categorial schemes, indeterminacy is structural and arises from partial overlap between categories. We study how the theory of belief functions can accurately describe the different sets of labels proposed by annotators, by explicitly representing uncertainty and disagreement.

Although non-univocal, this set-valued annotation preserves interpretative information and allows us to extract several annotated subcorpora and measure their quality. We also identify relevant groupings of labels that reduce uncertainty without losing semantic granularity.

The resulting annotated corpus is therefore accurate and independent of the learning tasks for which it is intended, while providing a framework that can improve model accuracy and refine model evaluation.

Keywords

Annotation NLP Uncertainty Belief functions

1 Introduction

Ce travail se situe à l'intersection de la théorie de l'annotation en TALN (méthodologique), de la modélisation formelle de l'incertitude (incertitude épistémique structurelle) et de l'étude des mécanismes décisionnels humains (décisions d'annotation) dans les processus de catégorisation (processus de décision sous indétermination interprétative). De nombreux travaux en annotation linguistique et fonctionnelle reposent sur l'hypothèse implicite d'une correspondance déterministe (univoque) entre une instance et une catégorie [4, 23]. Or, dans plusieurs contextes d'annotation fonctionnelle, les prédicats définissant les labels ne sont pas strictement disjoints : une même instance peut satisfaire simultanément plusieurs critères catégoriels valides.

Formellement, l'annotation est modélisée comme une fonction

$$f : I \to L,$$

où I désigne l'ensemble des instances annotées et L l'ensemble des étiquettes du schéma d'annotation [14]. Dans un contexte d'apprentissage supervisé, ces étiquettes correspondent aux labels associés aux instances du jeu de données [17]. Dans cet article, nous utilisons le terme *label* pour désigner les catégories du schéma d'annotation, afin de conserver une cohérence avec leur utilisation ultérieure comme variables cibles dans les jeux de données d'apprentissage supervisé.

Nous examinons un cas d'usage concret : l'annotation fonctionnelle d'interactions (ou prompts) avec des IA génératives, produites par des étudiants lors d'un hackathon. Le schéma d'annotation, inspiré de [13], repose sur 21 catégories fonctionnelles et une assignation mono-label.

Nous montrons que, malgré un protocole rigoureux, certaines interactions ne peuvent être assignées de manière univoque à un seul label.

Nous soutenons que cette situation ne relève pas nécessairement d'une erreur d'annotation ni d'un manque d'information, mais d'une propriété structurelle de l'espace catégoriel. L'annotation doit alors être modélisée non comme un problème de classification déterministe, mais comme un processus de décision sous incertitude interprétative, correspondant à une attribution d'incertitude structurée.

Plutôt que de modéliser l'assignation comme une fonction à valeur unique, nous proposons une fonction ensembliste

$$F : I \to \mathcal{P}(L),$$

où $\mathcal{P}(L)$ désigne l'ensemble des parties de L. L'annotation devient alors une relation entre interactions et sous-ensembles de labels compatibles.

Nous introduisons également, pour chaque instance annotée, un score dans l'intervalle $[0, 1]$, correspondant à la confiance de l'annotateur dans sa compréhension du prompt. Ce score ne porte donc pas sur la validité d'un label donné, mais sur l'intelligibilité de l'instance elle-même. Il permet de distinguer deux formes d'incertitude : d'une part, l'incertitude interprétative liée à une compréhension partielle ou fragile du prompt ; d'autre part, l'indétermination catégorielle, qui subsiste même lorsque le prompt est bien compris, du fait du recouvrement entre labels.

Cette reformulation permet de représenter explicitement la pluralité des labels valides pour une même instance. Nous introduisons la notion de *complexité interprétative*, définie pour une instance i par :

$$C(i) = |F(i)|\,,$$

afin de caractériser le degré d'indétermination catégorielle. Ce cadre formel est opérationnalisé au moyen des fonctions de croyance issues de la théorie de Dempster-Shafer [8, 24], qui permettent de représenter et de combiner cette incertitude sur les ensembles de labels.

1.1 Travaux relatifs

Trois axes de recherche sont directement pertinents pour notre approche : les limites des schémas d'annotation mono-label, la reconsidération du désaccord d'annotation, et les conséquences pour la modélisation en apprentissage automatique.

1.1.1 Annotation mono-label et ses limites

L'hypothèse d'un label unique par instance suppose implicitement que les prédicats définissant les catégories sont mutuellement exclusifs et exhaustifs. Or cette hypothèse ne tient pas dans tous les contextes.

Si la multifonctionnalité des énoncés a été reconnue théoriquement dès les premières formalisations des actes de dialogue [1, 5] et formalisée dans le standard ISO [6], la pratique d'annotation dominante a néanmoins maintenu l'assignation mono-label [4, 23]. Cette persistance s'explique par des contraintes opérationnelles, telles que le coût de l'annotation et la complexité de l'évaluation inter-annotateurs, plutôt que par une adhésion théorique à l'hypothèse déterministe.

1.1.2 "Désaccord annotatif" comme signal structurel

Dans des travaux de recherche récents, on trouve un courant qui remet en question l'interprétation du désaccord annotatif comme une erreur à éliminer [2, 27]. [2] identifie plusieurs mythes fondateurs de la pratique d'annotation, dont celui du *gold standard* unique, et proposent de traiter le désaccord comme une information à modéliser. [19] montre empiriquement que le désaccord sur des tâches d'inférence textuelle reflète des divergences interprétatives légitimes plutôt que des erreurs. [20] généralise ce constat à l'ensemble des pratiques d'annotation en NLP et plaide pour une révision profonde de la notion de vérité terrain.

Ces travaux partagent l'idée que le désaccord est informatif. Ils diffèrent de notre approche sur un point essentiel : ils traitent le désaccord comme un phénomène *inter-annotateurs*, lié à la variation des jugements humains.

Nous proposons ici une lecture complémentaire : l'indétermination peut être une propriété *intra-instance*, indépendante du nombre d'annotateurs ou de leurs divergences, et inhérente au recouvrement partiel des catégories du schéma d'annotation lui-même. Dans ce sens, le caractère *structurel* du désaccord ne renvoie pas à un manque d'information ni à une erreur d'annotation, mais à la structure de l'espace catégoriel.

1.1.3 Conséquences pour la modélisation en apprentissage automatique

Lorsque plusieurs labels peuvent être compatibles avec une même instance, le problème ne relève plus de la classification mono-label classique. La littérature en apprentissage automatique traite ces situations dans le cadre de la *multi-label classification*, où une instance peut être simultanément associée à plusieurs labels [26]. Dans la plupart des travaux existants, cette pluralité de labels est mise en évidence empiriquement à partir des données annotées, par exemple lorsque plusieurs annotateurs produisent des jugements divergents qui sont conservés sous forme de distributions ou d'ensembles de labels [2, 27]. Dans ce cas, la pluralité est interprétée comme résultante de la variation inter-annotateurs. Dans le contexte de certaines tâches d'annotation linguistique, plusieurs travaux ont également montré que le choix forcé entre catégories peut masquer une ambiguïté réelle et introduire un biais dans l'évaluation de la fiabilité annotative [21]. L'approche proposée ici repose sur une hypothèse différente : la pluralité des labels ne résulte pas d'un agrégat de jugements d'annotation mais d'une propriété structurelle de l'espace catégoriel.

1.2 Positionnement par rapport aux pratiques existantes

Dans la pratique courante de l'annotation fonctionnelle, le désaccord d'annotation est généralement interprété comme une erreur ou une divergence à résoudre, et évalué au moyen de mesures de fiabilité telles que le coefficient κ [7, 3]. Cette hypothèse d'un "gold label" unique fait cependant l'objet de discussions croissantes [2, 27, 20] : le désaccord entre annotateurs y est reconsidéré comme un signal potentiellement informatif plutôt que comme un bruit à éliminer. Ces travaux expliquent toutefois l'indétermination principalement par la variation des jugements humains. Dans ce cadre, l'objectif est de réduire l'indétermination.

Nous proposons ici une lecture complémentaire : l'indétermination peut également être une propriété *intra-instance*, révélant un recouvrement partiel des catégories du schéma d'annotation lui-même. Dans ce cas, la pluralité des labels ne doit pas être réduite à une décision exclusive mais peut être explicitement représentée et intégrée au processus annotatif. Cette observation conduit à reformuler l'annotation comme une relation entre instances et labels. Pour représenter et opérationnaliser cette incertitude structurelle, nous recourons à la théorie des fonctions de croyance [8, 24], qui offre un cadre formel permettant d'attribuer des degrés de confiance à des ensembles de labels plutôt qu'à des labels uniques.

2 Cadre d'annotation existant

2.1 Schéma d'annotation

L'annotation (type de catégorie) des interactions s'inspire du protocole proposé par Guner et al. [13] développé pour l'analyse des échanges étudiants-outils IA générative (ChatGPT, Gemini, etc.) dans un contexte d'apprentissage de la programmation. Dans ce schéma, chaque prompt constitue l'unité d'annotation. Les auteurs définissent 21 catégories fonctionnelles, visant à caractériser la *finalité* du prompt dans l'interaction. L'annotation repose sur une analyse pragmatique du contenu de la requête, centrée sur l'intention communicative exprimée par l'étudiant, indépendamment des résultats obtenus ou de la performance académique ultérieure. Les catégories couvrent un ensemble de fonctions interactionnelles comme la reproduction des instructions, la demande d'explication conceptuelle, la révision de code, la définition du contexte, etc. (cf. Table 1). Dans le protocole original, chaque prompt est assigné à une unique catégorie fonctionnelle.

L'opération de labellisation est réalisée par deux experts, puis discutée jusqu'à l'obtention d'un consensus. Cette procédure vise à garantir la cohérence interprétative et à limiter les biais individuels.

2.2 Adaptation au cas du hackathon

Dans notre étude, ce schéma d'annotation est appliqué à un corpus de prompts produits lors d'un hackathon étudiant dans le cadre d'un cours d'IA à l'université. L'annotation initiale repose sur une assignation mono-label : chaque prompt est annoté indépendamment par deux experts, puis les divergences éventuelles sont arbitrées par un troisième annotateur afin d'aboutir à un label unique.

Toutefois, au cours du processus d'annotation, nous avons observé plusieurs cas où plusieurs labels apparaissaient simultanément compatibles avec une même instance. Ces situations ne résultaient pas d'un manque d'information ni d'un désaccord interprétatif entre les annotateurs, mais de la compatibilité sémantique entre plusieurs catégories du schéma.

Afin d'explorer cette indétermination, nous avons également réalisé une seconde annotation autorisant l'assignation de plusieurs labels par prompt, accompagnée d'un degré de confiance pour chaque attribution.

Dans ce travail, l'analyse porte exclusivement sur le contenu des prompts eux-mêmes : les interactions sont considérées indépendamment de leur contexte conversationnel ou du profil des étudiants.

3 Indétermination structurelle

Dans ce travail, nous considérons que l'annotation d'une interaction $i \in I$ peut être associée non pas à un label unique, mais à un ensemble de labels plausibles $F(i) \subseteq L$, tel qu'introduit précédemment. Cette représentation permet de rendre explicites les situations dans lesquelles plusieurs catégories apparaissent simultanément compatibles avec une même instance.

Même avec une guideline claire, une consigne linguistique stricte et l'expertise des annotateurs, certaines interactions peuvent être interprétées de plusieurs façons fonctionnelles valides.

3.1 Non-orthogonalité des catégories

Les catégories définies dans le schéma d'annotation couvrent un ensemble de fonctions interactionnelles relativement fines. Cependant, certaines d'entre elles ne sont pas mutuellement exclusives sur le plan interprétatif.

Un même prompt peut ainsi relever simultanément de plusieurs catégories fonctionnelles. Par exemple, une interaction peut être à la fois une *Coding task inquiry*, une *Concept explanation* et une *Defining context*. Cette superposition résulte du fait que l'espace catégoriel n'est pas orthogonal : les prédicats définissant les labels se recouvrent partiellement. L'indétermination observée ne provient pas nécessairement d'une divergence entre annotateurs, mais de la structure du schéma d'annotation.

3.2 Complexité interprétative

Dans ce cadre, il devient possible de caractériser le degré d'indétermination associé à une interaction.

Prenons l'exemple :

> "Is this correct ?"

Cet énoncé peut correspondre à une vérification, à une demande implicite d'explication ou à une validation finale. Nous définissons la complexité interprétative d'une interaction i comme :

$$C(i) = |F(i)|$$

Type de prompt	Explication de prompt
1. Exact copy of instruction	Directly copy-pasting entire task instructions.
2. Partial copy of instruction	Partially copy-pasting task instructions.
3. Summary of instruction	Summarizing or paraphrasing task instructions.
4. Coding task inquiry	Asking for specific coding knowledge or guidance related to a particular task or requirement.
5. Refining GPT code	Asking for refinement on the GPT solution to meet the task requirements.
6. Refining own code	Asking for refinement on their own code to meet the task requirements.
7. Debugging own code	Asking for finding errors in their own code.
8. Debugging GPT code	Asking for fixing errors in GPT solution.
9. Reviewing own code	Requesting review, enhancement, or optimization for their own code.
10. Reviewing the GPT code	Requesting review, enhancement, or optimization for GPT code.
11. GPT code explanation	Asking for explanation about GPT solution.
12. Prompt revision	Requesting GPT to modify its previous response based on feedback or clarification.
13. Concept explanation	Asking for explanation on related concepts.
14. Asking for more info	Asking for clarification or additional information about a given response.
15. Generating alternative code	Asking GPT to simplify its previous solution or to offer another alternative solution.
16. Defining context	Establishing the context, scope, or requirements before the main prompt.
17. Combining all GPT code	Asking to combine previous GPT responses into one.
18. Code for given output	Asking GPT to code for a given output.
19. Verifying GPT solution	Asking GPT for verification of its solution.
20. Suggested inquiry	Suggested prompts from the guide document.
21. Other	Typos, spelling errors, obscure statements, greetings, or gratitude expressions.

TABLE 1 – Catégories fonctionnelles (des prompts) et leurs définitions ([13]).

où $F(i)$ désigne l'ensemble des labels plausibles. Lorsque $C(i) = 1$, l'assignation est univoque. Lorsque $C(i) > 1$, l'interaction présente une indétermination structurelle que le schéma mono-label force artificiellement à réduire. Cette pluralité d'interprétations ne relève pas d'une ambiguïté lexicale mais de la coexistence de plusieurs dimensions fonctionnelles au sein d'un même énoncé et doit pouvoir être représentée explicitement dans le modèle d'annotation.

4 Formalisation dans le cadre des fonctions de croyance

4.1 Théorie des fonctions de croyance

L'interprétation "probale" (*evidential*) de la théorie des fonctions de croyance [8, 24] permet de modéliser les degrés de confiance issus des annotations ensemblistes effectuées par les experts. Introduisons d'abord quelques définitions. Soit $i \in I$ une interaction. Le *cadre de discernement* est l'ensemble des labels $L = \{1, \ldots, n\}$ pouvant décrire i. Une fonction de masse $m_i : \mathcal{P}(L) \to [0, 1]$ vérifie $m_i(\emptyset) = 0$ et $\sum_{B \subseteq L} m_i(B) = 1$. Chaque ensemble $B \subseteq L$ tel que $m_i(B) > 0$ est appelé un "élément focal de m_i" et représente le fait que l'on dispose de l'information $F_i = B$ avec un degré de confiance égal à $m(B) \in [0, 1]$. Intuitivement, B est l'ensemble de labels donné par un expert.

À partir de la fonction de masse m_i, on peut définir deux mesures, Bel_i et $\mathrm{Pl}_i : \mathcal{P}(L) \to [0, 1]$, appelées respectivement fonctions de croyance et de plausibilité, définies par :

$$\mathrm{Bel}_i(A) = \sum_{B \subseteq A} m_i(B) \quad \text{et} \quad \mathrm{Pl}_i(A) = \sum_{B \cap A \neq \emptyset} m_i(B)$$

Ces mesures sont duales : $\mathrm{Bel}_i(A) + \mathrm{Pl}_i(\bar{A}) = 1$ pour tout $A \subseteq L$. $\mathrm{Bel}_i(A)$ indique à quel point l'ensemble d'étiquettes $A \subseteq L$ est impliqué par les annotations des experts, et $\mathrm{Pl}_i(A)$ indique à quel point il est compatible (cf. exemple 1). Notons que nous avons deux familles de cas particuliers : si tous les experts font des annotations précises, alors les éléments focaux sont des singletons ($\forall B$ focal, $|B| = 1$) et $\mathrm{Bel}_i = \mathrm{Pl}_i$ est une mesure de probabilité sur L ; par ailleurs si tous les experts font des annotations compatibles entre elles, alors les éléments focaux sont emboîtés ($\forall B_1$ et B_2 focaux, $B_1 \subseteq B_2$ ou $B_2 \subseteq B_1$) et Pl_i est une mesure de possibilité et Bel_i est sa mesure de nécessité duale [12].

Example 1 (Annotation) *Considérons l'annotation de l'interaction i, avec $L = \{1, 2, 3\}$.*

- *L'expert a dit que $F_a(i) = \{1, 2\}$.*
- *L'expert b dit que $F_b(i) = \{2, 3\}$.*

On accorde autant de confiance à chaque expert, ce qu'on modélise par la fonction de masse m_i telle que :

$$m_i(\{1, 2\}) = m_i(\{2, 3\}) = 0.5 \qquad m_i(B) = 0 \text{ sinon.}$$

Comme on l'observe dans la table 2, la fonction de croyance (resp. de plausibilité) issue de m_i indique le degré d'implication (resp. de compatibilité) avec les annotations effectuées par les experts, pour chaque label et ensemble de labels. Ainsi, par exemple, le label 2 est complètement compatible ($\mathrm{Pl}_i(\{2\}) = 1$). L'ensemble de labels $\{1, 2\}$ est impliqué à 50% ($\mathrm{Bel}(\{1, 2\} = 0.5$) car impliqué par l'expert a uniquement. Notons que $\emptyset$ est nécessairement incompatible (contradiction), et que L est nécessairement certain (tautologie).

A	$m_i(A)$	$\mathrm{Bel}_i(A)$	$\mathrm{Pl}_i(A)$
$\emptyset$		0	0
$\{1\}$		0	0.5
$\{2\}$		0	1
$\{3\}$		0	0.5
$\{1,2\}$	0.5	0.5	1
$\{1,3\}$		0	1
$\{2,3\}$	0.5	0.5	1
$\{1,2,3\}$		1	1

TABLE 2 – Masses et degrés de croyance et de plausibilité pour l'exemple 1

Interprétation des valeurs. Dans l'interprétation probale, on suppose habituellement qu'un état (ici, une étiquette) $l^* \in L$ est le « vrai », que tout ensemble $A \subseteq L$ décrit la proposition « $l^* \in A$ » et que les valeurs $\mathrm{Bel}(A)$ et $\mathrm{Pl}(A)$ décrivent la croyance en cette proposition. Or, ici, nous nous intéressons à l'incertitude provenant de l'indétermination structurelle, et la « vérité » est un ensemble $A^* \subseteq L$ contenant toutes les étiquettes correctes. Ainsi, tout ensemble A représente la proposition « A contient l'une des étiquettes correctes » (i.e., $A \cap A^* \neq \emptyset$), et $\mathrm{Bel}(A)$ (resp. $\mathrm{Pl}(A)$) indique à quel point cette proposition est impliquée (resp. compatible) avec la connaissance.
Notons que cette interprétation diffère de l'interprétation en termes de famille de probabilités (ou *ensemble crédal*). Dans l'exemple 1, $\mathrm{Bel}(\{1,2\}) = 0.5$ et $\mathrm{Pl}(\{1,2\}) = 1$ indiquent que la moitié des annotations implique les étiquettes $\{1,2\}$ et que la totalité des annotations sont compatible avec ces étiquettes. Cependant, cela ne signifie généralement pas que la probabilité que l'étiquette correcte soit 1 ou 2 est comprise entre 0.5 et 1. En effet, l'intérêt du cadre proposé est de gérer l'incertitude structurelle, où une interaction ne possède pas forcément d'étiquette unique : une interaction peut avoir plusieurs étiquettes valides simultanément (par exemple, {1,2} ou {2,3}), ce qui n'est pas représentable par une distribution de probabilité.

4.2 Protocole d'annotation

On définit un protocole d'annotation (Algorithme 1) dans lequel l'expert peut attribuer à chaque instance soit un label unique, soit un ensemble de labels plausibles. À chaque annotation est associé un degré de confiance $d_i^x \in [0,1]$, reflétant le niveau de confiance de l'expert x dans la compréhension du prompt i, c'est-à-dire la clarté et la non-ambiguïté de l'énoncé. Le degré de confiance dans la compréhension du prompt est indépendant du nombre de labels attribués. Autrement dit, l'annotateur peut être très sûr de sa compréhension (degré proche de 1) tout en considérant que plusieurs catégories sont simultanément pertinentes.
Il serait également possible de donner un degré de confiance pour chacun des labels et/ou sous-ensembles des labels choisis, afin d'exprimer plus finement la certitude de l'expert. Toutefois, l'introduction de cette dimension compliquerait la définition du degré de confiance, rendrait la tâche d'annotation fastidieuse et entraînerait une plus grande complexité ; nous choisissons donc de ne pas la prendre en compte dans ce travail : chaque annotateur donne un unique ensemble d'étiquettes et un unique degré de confiance.

Algorithm 1 Protocole d'annotation multi-experts

Entrées :
— Ensemble d'instances $I = \{i_1, \ldots, i_n\}$
— Ensemble de labels $L = \{l_1, \ldots, l_m\}$
— Ensemble d'experts $X = \{x_1, \ldots, x_p\}$
Sortie : Ensemble d'annotations A
Pour chaque instance $i \in I$ **faire**
  Présenter le prompt correspondant à l'instance
  Pour chaque expert $x \in X$ **faire**
    L'expert x interprète le prompt
    L'expert x sélectionne l'annotation B_i^x telle que $B_i^x \in L$ ou $B_i^x \subseteq L$, $|B_i| \geq 1$
    L'expert x attribue un degré de confiance :
      $d_i^x = f(\text{compréhension du prompt}) \in [0,1]$
    Enregistrer l'annotation : $a_i^x = (x, B_i^x, d_i^x)$
    Ajouter a_i^x à A
  Fin Pour
Fin Pour
Return A

4.3 Regroupement des annotations

Nous étudions ici différentes façons de regrouper les annotations effectuées par les experts : par *pooling*, par combinaison conjonctive ou par combinaison disjonctive. Rappelons d'abord que chaque annotateur $x \in X$ indique, pour une interaction $i \in I$ donnée, l'ensemble B_i^x des étiquettes qui lui semble pertinent ainsi que son degré de certitude, noté d_i^x.

Pooling. La méthode la plus directe nous semble être le *pooling* des annotations effectuées. Cela correspond à l'interprétation probale telle que décrite dans le modèle à croyance transférables [25] : chaque expert produit un élément de preuve (*piece of evidence*), et la connaissance jointe est simplement l'ensemble des éléments de preuve. La fonction de masse décrivant la connaissance jointe sur l'interaction $i \in I$ est m_i, définie par :

$$m_i(B) = \frac{1}{K_i} \sum_{x \in X, B = B_i^x} d_i^x \quad \text{et} \quad m_i(B) = 0 \text{ sinon.}$$

Les masses sont proportionnelles aux degrés de confiance des experts ; $K_i = \sum_x d_i^x$ est le facteur de normalisation.
C'est la méthode utilisée dans l'exemple 1 (avec des degrés de certitude égaux : $d_i^a = d_a^b > 0$).
Le *pooling* offre ainsi une interprétation claire, tout en gardant la tâche d'annotation simple. Notons qu'à cause de la normalisation, les degrés de certitudes sont considérés de manière relative uniquement. De plus, il n'est pas possible d'itérer le *pooling* (si les annotations d'un autre expert doivent être ajoutées ultérieurement à m_i), car il faut connaître la valeur K_i utilisée afin de re-normaliser. Cette limitation peut être dépassée en conservant cette valeur, ou en utilisant le cadre proposé dans [22] (qui permettrait aussi

d'exprimer le doute sur la fiabilité des experts, ce que nous n'étudions pas ici).
Les éléments focaux de m_i sont les ensembles indiqués par les annotateurs : $\mathcal{S}(m_i) = \{B_i^x \mid x \in X\}$; la complexité est linéaire en le nombre d'annotateurs.

Combinaison conjonctive. La règle de combinaison conjonctive $\otimes_\cap$ [24] et la règle de Dempster $\otimes_\oplus$ [8] permettent de combiner des informations exprimées par des fonctions de masse.

$$m^\cap(B) = (m_1 \otimes_\cap m_2)(B) = \sum_{B_1 \cap B_2 = B} m_1(B_1) \times m_2(B_2),$$

$$m^\oplus(B) = (m_1 \otimes_\oplus m_2)(B) = \begin{cases} 0 & \text{si } B = \emptyset \\ \frac{m_{12}^\cap(B)}{1 - m_{12}^\cap(\emptyset)} & \text{sinon.} \end{cases}$$

La valeur $m^\cap(\emptyset)$ indique le degré de conflit entre m_1 et m_2, et la règle de Dempster ignore ce conflit (les valeurs sont donc renormalisées).
Nous pouvons appliquer ces règles pour combiner les annotations effectuées par les experts sur une même interaction. Nous modélisons alors chaque annotation par une *simple support function* [24] : l'annotation de l'interaction $i \in I$ par l'expert $x \in X$ est la fonction de masse m_i^x telle que :

$$m_i^x(B_i^x) = d_i^x \quad m_i^x(L) = 1 - d_i^x \quad m_i^x(B) = 0 \text{ sinon.}$$

Cela décrit le fait que l'expert x croit à l'ensemble B_i^x avec un degré d_i^x et ne sait pas sinon. Alors, la connaissance jointe concernant l'interaction $i \in I$ est la fonction de masse $m_i = m_i^{x_1} \otimes_\cap m_i^{x_2} \otimes_\cap \cdots \otimes_\cap m_i^{x_k}$.
Notons que si $\forall x, d_i^x = 1$, alors la combinaison conjonctive revient à identifier l'intersection des ensembles d'étiquettes donnés par les experts.
Supposons $d_i^x < 1$ pour tout $x \in X$. Alors, chaque fonction de masses m_i^x a deux éléments focaux : $\mathcal{S}(m_i^x) = \{B_i^x, L\}$. La complexité croit par combinaison : $\mathcal{S}(m_i^x \otimes_\cap m_i^y) = \{B_i^x \cap B_i^y, B_i^x, B_i^y, L\}$. Généralement, $\forall m, \mathcal{S}(m \otimes \cap m_i^x) = \{B \cap B_i^x \mid B \in \mathcal{S}(m)\} \cup \mathcal{S}(m)$, d'où $|\mathcal{S}(m \otimes \cap m_i^x)| \leq 2 \times |\mathcal{S}(m)|$. Par récurrence, nous avons $|\mathcal{S}(m_i^1 \otimes_\cap \cdots \otimes_\cap m_i^k)| \leq 2^k$, la complexité est exponentielle en le nombre d'annotateurs. Cela pourrait être prohibitif mais reste réaliste dans notre contexte où $k = 3$.

Combinaison disjonctive. [24] La combinaison disjonctive considère l'union des ensembles plutôt que leur intersection. La procédure est similaire : les annotations des experts sont exprimées par des *simple support functions* puis combinées par :

$$m^\cup(B) = (m_1 \otimes_\cup m_2)(B) = \sum_{B_1 \cup B_2 = B} m_1(B_1) \times m_2(B_2).$$

Aucun conflit ne peut apparaître avec la règle conjonctive. Notons que si $\forall x, d_i^x = 1$, alors la combinaison disjonctive revient à identifier l'union des ensembles d'étiquettes donnés par les experts.
Supposons $d_i^x < 1$ pour tout $x \in X$. Alors $\mathcal{S}(m_i^x) = \{B_i^x, L\}$. Ainsi, $\mathcal{S}(m_i^x \otimes_\cup m_i^y) = \{B_i^x \cup B_i^y, L\}$. Par récurrence, nous avons $\mathcal{S}(m_i^1 \otimes_\cup \cdots \otimes_\cup m_i^k) = \{\bigcup_{x=1}^k B_i^x, L\}$, d'où $|\mathcal{S}(m_i^1 \otimes_\cup \cdots \otimes_\cup m_i^k)| \leq 2$, la complexité est constante.

Impact de la méthode de regroupement. Les trois méthodes de regroupement étudiées décrivent des processus différents. La méthode « idéale » dépend de la façon dont les experts ont effectué l'annotation.
Le *pooling* est la méthode la plus directe, il ne fait pas d'hypothèse sur le comportement des experts. Elle regroupe simplement leurs annotations en un *pool* d'annotations. Notons que :

- les degrés de confiance éventuellement indiqués par les experts n'y sont considérés que de manière relative,
- les ensembles A tels que $\text{Bel}_i^{\text{Pool}}(A) = 1$ sont les ensembles qui incluent tous les labels identifiés par les experts,
- les ensembles A tels que $\text{Pl}_i^{\text{Pool}}(A) = 1$ sont les ensembles qui contiennent au moins un label identifié par chacun des experts.

En se focalisant sur un sous-ensemble de labels de croyance élevée et de taille minimale, les règles de combinaison peuvent être interprétées comme la première étape d'amélioration des étiquetages : la fusion des annotations revient effectivement à créer une nouvelle annotation jointe.

- La règle conjonctive mène à l'intersection des annotations, elle est adaptée si les experts ont cherché à être exhaustifs (quitte à donner des labels superflus).
- La règle disjonctive mène à l'union des annotations, elle est adaptée si les annotateurs ont cherché à être prudents (quitte à oublier certains labels pertinents).

Nous illustrons l'impact de ces trois méthodes de regroupement des annotations sur deux exemples.

Example 2 (Annotations compatibles) *Considérons l'annotation de l'interaction i avec 3 étiquettes possibles : $L = \{1, 2, 3\}$.*

- *L'expert a dit que $B_i^a = \{1\}$ avec un degré de confiance $d_i^a = 0.6$.*
- *L'expert b dit que $B_i^b = \{1, 2\}$ avec un degré de confiance $d_i^b = 0.9$.*

Le regroupement par *pooling* est direct, associant une masse normalisée de $\frac{d_i^x}{1.5}$ à chaque ensemble B_i^x. Ces masses sont proportionnelles aux degrés de certitude des experts, et mènent aux valeurs listées dans la table 3.
Pour les règles de combinaisons conjonctives et disjonctives, il faut d'abord exprimer les fonctions de masse m_i^x pour chaque expert $x \in X$.

$$m_i^a(\{1\}) = 0.6 \quad m_i^a(L) = 0.4 \quad m_i^a(B) = 0 \text{ sinon,}$$
$$m_i^b(\{1, 2\}) = 0.9 \quad m_i^b(L) = 0.1 \quad m_i^b(B) = 0 \text{ sinon.}$$

Les règles de combinaison disjonctive et conjonctive mènent aux valeurs listées dans la table 3. Puisqu'il n'y a pas de conflit ($m_i^\cap(\emptyset) = 0$), nous avons $m_i^\cap = m_i^\oplus$.

Example 3 (Annotations contradictoires) *Considérons l'annotation de l'interaction j avec 3 étiquettes possibles : $L = \{1, 2, 3\}$.*

- *L'expert a dit que $B_i^a = \{1\}$ avec un degré de confiance $d_i^a = 0.6$.*

A	$m_i^{\text{Pool}}(A)$	$\text{Bel}_i^{\text{Pool}}(A)$	$\text{Pl}_i^{\text{Pool}}(A)$	$m_i^{\cap}(A)$	$\text{Bel}_i^{\cap}(A)$	$\text{Pl}_i^{\cap}(A)$	$m_i^{\cup}(A)$	$\text{Bel}_i^{\cup}(A)$	$\text{Pl}_i^{\cup}(A)$
$\emptyset$		0	0		0	0		0	0
$\{1\}$	**0.6**/1.5	0.4	1	0.6	0.6	1		0	1
$\{2\}$		0	0.6		0	0.4		0	1
$\{3\}$		0	0		0	0.04		0	0.46
$\{1,2\}$	**0.9**/1.5	1	1	0.36	0.96	1	0.54	0.54	1
$\{1,3\}$		0.4	1		0.6	1		0	1
$\{2,3\}$		0	0.6		0	1		0	1
$\{1,2,3\}$		1	1	0.04	1	1	0.46	1	1

TABLE 3 – Masses normalisées et degrés de croyance et de plausibilité obtenus par pooling, par fusion disjonctive et par fusion conjonctive pour l'exemple 2.

A	$m_j^{\text{Pool}}(A)$	$\text{Bel}_j^{\text{Pool}}(A)$	$\text{Pl}_j^{\text{Pool}}(A)$	$m_j^{\mathfrak{D}}(A)$	$\text{Bel}_j^{\mathfrak{D}}(A)$	$\text{Pl}_j^{\mathfrak{D}}(A)$	$m_j^{\cup}(A)$	$\text{Bel}_j^{\cup}(A)$	$\text{Pl}_j^{\cup}(A)$
$\emptyset$		0	0		0	0		0	0
$\{1\}$	**0.6**/1.5	0.4	0.4	0.13	0.13	0.22		0	1
$\{2\}$		0	0.6	0.78	0.78	0.87		0	1
$\{3\}$		0	0.6		0	0.09		0	1
$\{1,2\}$		0.4	1		0.91	1		0	1
$\{1,3\}$		0.4	1		0.13	0.22		0	1
$\{2,3\}$	**0.9**/1.5	0.6	0.6		0.78	0.87		0	1
$\{1,2,3\}$		1	1	0.09	1	1	1	1	1

TABLE 4 – Masses normalisées et degrés de croyance et de plausibilité obtenus par pooling, par fusion de Dempster et par fusion conjonctive pour l'exemple 3.

— *L'expert b dit que* $B_i^b = \{2,3\}$ *avec un degré de confiance* $d_i^b = 0.9$.

Le regroupement par pooling est direct, associant une masse de $d_j^x/1.5$ à chaque ensemble B_j^x, ce qui mène aux valeurs présentées dans la table 4.

Pour les règles de combinaison, il faut d'abord exprimer les fonctions de masse m_j^x pour chaque expert $x \in X$.

$$m_j^a(\{1\}) = 0.6 \quad m_j^a(L) = 0.4 \quad m_j^a(B) = 0 \text{ sinon,}$$
$$m_j^b(\{2\}) = 0.9 \quad m_j^b(L) = 0.1 \quad m_j^b(B) = 0 \text{ sinon.}$$

La règle de combinaison disjonctive produit ici un conflit : $m_j^{\cap}(\emptyset) = 0.54 > 0$. La suppression du conflit et renormalisation mènent à la règle de Dempster. Les résultats des règles de Dempster et conjonctives sont listés dans la table 4.

4.4 Qualité des annotations

De nombreuses métriques ont été proposées dans la littérature pour quantifier l'incertitude présente dans une fonction de croyance. Ces métriques visent généralement à mesurer l'imprécision, le conflit ou une combinaison de ces deux dimensions.

Imprécision. Les métriques mesurant l'imprécision se basent sur la cardinalité des éléments focaux (spécificité, entropie de croyance [28, 11, 10, 9]). Dans notre contexte, cela représente l'incertitude structurelle liée aux prompts, qui peuvent être associés à plusieurs étiquettes plausibles.

Nous considérons ici la mesure de spécificité [28] :

$$Sp(m) = \sum_{A \subseteq L} \frac{m(A)}{|A|}.$$

La valeur maximale $Sp(m) = 1$ est atteinte lorsque les annotations de tous les experts sont des singletons (considérant la combinaison disjonctive, il faut en plus que tous les experts soient d'accord sur le même singleton). La valeur minimale $Sp(m) = 1/|\Omega|$ est obtenue lorsque tous les experts affirment conjointement que toutes les étiquettes sont compatibles ($m(L) = 1$).

Conflit. Les métriques mesurant le conflit se basent sur la disjonction des éléments focaux (discorde, dissonance, conflit [16, 29, 15]). Dans notre contexte, elles indiquent donc le désaccord entre les experts.

Nous considérons ici la mesure de dissonance [15] :

$$D(m) = -\sum_{A \subseteq L} m(A) \log_2 \left(1 - \sum_{B \subseteq L} m(B) \frac{|A \cap B|}{|A|}\right)$$

Le terme $\frac{|A \cap B|}{|A|}$ mesure le degré de compatibilité entre deux ensembles de labels, et représente le degré d'accord entre les experts.

Utilisation de ces métriques. Ces deux métriques permettent d'évaluer la qualité des annotations pour chaque interaction. Nous pouvons envisager principalement deux usages :

— **Identifier les annotations problématiques.** Si l'annotation d'une interaction $i \in I$ est décrite

par une fonction de masse fortement dissonante ($D(m_i) > x$, où x est un seuil fixé), cela indique une forte contradiction entre les experts. L'interaction doit alors être réexaminée dans un processus itératif.

— **Évaluer la qualité globale du corpus.** La moyenne des spécificités ou des dissonances permet d'estimer la qualité globale du jeu de données du point de vue de l'incertitude structurelle ou du désaccord entre experts, respectivement.

La capacité à évaluer un jeu de données selon ces deux critères permet non seulement d'analyser l'incertitude présente dans les annotations, mais aussi d'exploiter ces informations pour construire différentes versions du corpus annoté.

5 Sélection de sous-corpus de qualité

Plusieurs stratégies peuvent être envisagées pour produire de nouveaux sous-corpus à partir du jeu de données annoté.

Correction des annotations. Une première approche consiste à corriger les annotations afin de réduire les situations de conflit ou d'imprécision. Toutefois, cette stratégie est en tension avec l'objectif initial de notre protocole, qui vise précisément à préserver l'information portée par l'indétermination annotative. Elle pourrait néanmoins être envisagée dans des cas particuliers, par exemple lorsque la grande majorité des experts s'accorde sur un même ensemble de labels.

Filtrage des interactions. Une seconde stratégie consiste à ne conserver qu'un sous-ensemble des interactions selon des critères de qualité. Par exemple, on peut sélectionner uniquement les instances pour lesquelles la dissonance est inférieure à un seuil donné ou celles dont les annotations présentent un degré élevé de spécificité.

Une telle sélection peut poursuivre deux objectifs : améliorer la cohérence globale du jeu de données utilisé pour l'entraînement ou l'évaluation des modèles, ou isoler des sous-ensembles présentant certaines propriétés particulières, par exemple les interactions dont l'annotation est univoque.

Regroupement des catégories. Une troisième stratégie consiste à regrouper certaines catégories de labels afin de réduire le recouvrement observé dans l'espace catégoriel. Cette opération peut conduire à définir des catégories plus larges pour lesquelles chaque interaction est associée à une étiquette unique.

Ces différentes transformations permettent de générer plusieurs versions du jeu de données, adaptées à différents objectifs d'analyse ou d'apprentissage.

6 Application du cadre proposé au jeu de données « Hackaton »

L'un des intérêts des annotations ensemblistes est qu'elles ne se contentent pas de signaler une indétermination locale : elles permettent aussi de mettre en évidence la structure effective de l'espace catégoriel à partir des combinaisons de labels observées.

Identification des catégories. On observe que les labels 5, 6, 7, 8, 9 et 10 de la table 1 sont souvent regroupés. On trouve par exemple :

— Des interactions pour lesquelles on observe les annotations $\{5, 6\}$, $\{7, 8\}$ ou $\{9, 10\}$ (il s'agit clairement de "*refining*", de "*debugging*" ou de "*reviewing*", mais le code est un mélange de "*own code*" et de "*GPT code*").

— Des interactions pour lesquelles on observe les annotations $\{5, 7, 9\}$ ou $\{6, 8, 10\}$ (il s'agit clairement de "*own code*" ou de "*gpt code*", mais la tache est un mélange de "*refining*", "*debugging*" et "*reviewing*").

— Des interactions pour lesquelles on observe l'annotation $\{5, 6, 7, 8, 9, 10\}$: il s'agit d'une tache mixte pour un code mixte.

La structure de l'espace catégoriel mise en évidence n'est pas une hiérarchie, mais semble s'organiser en treillis (en effet, $\{5\} \subset \{5, 6\} \subset \{5, 6, 7, 8, 9, 10\}$ et $\{5\} \subset \{5, 7, 9\} \subset \{5, 6, 7, 8, 9, 10\}$, mais $\{5, 6\} \not\subseteq \{5, 7, 9\}$ par exemple).

Cette approche permet, en particulier, d'identifier des regroupements d'étiquettes mutuellement exclusifs (tels que la spécificité des annotations de chaque interaction soit égale à 1). Ici, par exemple, les labels $\{5, 6, 7, 8, 9, 10\}$ peuvent être regroupés en une catégorie "code". De cette manière, on construit un jeu de données annoté de manière univoque, en résolvant l'incertitude structurelle identifiée par les experts de manière exacte et sans introduire d'hypothèse arbitraire supplémentaire.

7 Discussion et perspectives

Implications pour l'annotation. Le consensus annotatif ne supprime pas nécessairement l'indétermination structurelle ; il peut au contraire la masquer en imposant une décision exclusive là où plusieurs interprétations demeurent simultanément compatibles. L'annotation mono-label apparaît ainsi moins comme une propriété intrinsèque des données que comme une contrainte méthodologique imposée par le schéma d'annotation.

Perspectives. Plusieurs prolongements de ce travail peuvent être envisagés. D'une part, l'analyse systématique des cooccurrences de labels pourrait permettre de mieux caractériser la structure de l'espace catégoriel et de proposer des schémas d'annotation agrégés. D'autre part, l'intégration du contexte interactionnel constitue une piste importante. Dans ce travail, les prompts ont été annotés de manière isolée afin de mettre en évidence l'indétermination structurelle liée au schéma catégoriel. Cependant, dans certaines situations, l'accès au contexte de l'interaction (prompts précédents, état du code et progression de la tâche) pourrait aider l'annotateur à mieux interpréter la fonction d'un prompt et à préciser l'ensemble des labels plausibles. L'intégration du contexte interactionnel dans le processus d'annotation permettrait de constituer des jeux de données reliant les fonctions des prompts à la progression du travail de l'étudiant et à l'évolution de ses

stratégies d'usage de l'IA.

8 Conclusion

Nous avons montré que, dans le cas étudié, l'annotation fonctionnelle ne peut être modélisée de manière adéquate par une fonction déterministe associant chaque instance à un unique label. L'indétermination observée est structurelle et résulte du recouvrement partiel des catégories du schéma d'annotation. Nous proposons un protocole d'annotation multi-experts permettant une annotation multilabel, dans lequel chaque expert associe à une instance un ensemble de labels plausibles, accompagné d'un degré de confiance reflétant sa certitude quant à l'interprétation du prompt. Pour formaliser cette incertitude épistémique, nous utilisons le formalisme des fonctions de croyance, qui permet de représenter les ensembles de labels proposés par les annotateurs et de quantifier la qualité des annotations en fonction de leur imprécision ou du degré de conflit entre les experts.
Cette approche permet notamment d'identifier des sous-corpus annotés selon différents niveaux de qualité et de mettre en évidence des regroupements pertinents de catégories, révélant la structure effective de l'espace catégoriel. Enfin, en s'inspirant des approches de classification imprécise [18], notre protocole permet de produire des annotations ensemblistes ou crédales, reflétant explicitement l'incertitude associée aux labels des instances. Ces annotations ouvrent la voie à des décisions robustes fondées sur la mesure de l'incertitude, telles que la sélection d'instances ambiguës, la prédiction avec option de rejet ou la génération de prédictions à valeurs multiples.

Références

[1] James Allen and Mark Core. Draft of DAMSL : Dialog act markup in several layers. Technical report, University of Rochester, 1997. Version 2.1.

[2] Lora Aroyo and Chris Welty. Truth is a lie : Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1) :15–24, 2015.

[3] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4) :555–596, 2008.

[4] Petra Saskia Bayerl and Karsten Ingmar Paul. What determines inter-coder agreement in manual annotations ? a meta-analytic investigation. *Computational Linguistics*, 37(4) :699–725, 2011.

[5] Harry Bunt. Multifunctionality in dialogue. *Computer Speech & Language*, 25(2) :222–245, 2011. Language and speech issues in the engineering of companionable dialogue systems.

[6] Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 549–558, Marseille, France, May 2020. European Language Resources Association.

[7] Jean Carletta. Assessing agreement on classification tasks : The kappa statistic. *Computational Linguistics*, 22(2) :249–254, 1996.

[8] Arthur P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2) :325–339, 1967.

[9] Yong Deng. Deng entropy. *Chaos, Solitons & Fractals*, 91 :549–553, 2016.

[10] Didier Dubois and Henri Prade. A note on measures of specificity for fuzzy sets. *International Journal of General System*, 10(4) :279–283, 1985.

[11] Didier Dubois and Henri Prade. The principle of minimum specificity as a basis for evidential reasoning. *Bulletin pour les sous-ensembles flous et leurs applications*, 25 :124–132, 1985.

[12] Didier Dubois and Henri Prade. *Possibility Theory*. Springer US, Boston, MA, 1988.

[13] Hacer Güner and Erkan Er. Ai in the classroom : Exploring students' interaction with chatgpt in programming learning. *Education and Information Technologies*, pages 1–27, 2025.

[14] Nancy Ide and Jean Véronis, editors. *Corpus Annotation : Linguistic Information from Computer Text Corpora*. Kluwer Academic Publishers, Dordrecht, 1998.

[15] George J Klir and Behzad Parviz. A note on the measure of discord. In *Uncertainty in Artificial Intelligence*, pages 138–141. Elsevier, 1992.

[16] George J Klir and Arthur Ramer. Uncertainty in the dempster-shafer theory : a critical re-examination. *International Journal of General System*, 18(2) :155–166, 1990.

[17] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

[18] Vu-Linh Nguyen, Haifei Zhang, and Sébastien Destercke. Credal ensembling in multi-class classification. *Machine Learning*, 114(1) :19, 2025.

[19] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7 :677–694, 2019.

[20] Barbara Plank. The "problem" of human label variation : On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[21] Massimo Poesio and Ron Artstein. The reliability of anaphoric annotation, reconsidered : taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II : Pie in the Sky*, CorpusAnno '05, page 76–83, USA, 2005. Association for Computational Linguistics.

[22] Pierre Pomeret-Coquot. Modeling and updating uncertain evidence within belief function theory. *International Journal of Approximate Reasoning*, 182 :109428, 2025.

[23] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Sebastopol, CA, 2012.

[24] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[25] Philippe Smets and Robert Kennes. The transferable belief model. *Artificial intelligence*, 66(2) :191–234, 1994.

[26] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification : An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3) :1–13, July 2007.

[27] Alexandra Uma, Barbara Plank, Dirk Hovy, Tommaso Fornaciari, and Massimo Poesio. Learning from disagreement : A survey. *Journal of Artificial Intelligence Research*, 72 :1385–1470, 2021.

[28] Ronald R Yager. Entropy and specificity in a mathematical theory of evidence. *International Journal of General System*, 9(4) :249–260, 1983.

[29] Ronald R Yager. On the Dempster-Shafer framework and new combination rules. *Information sciences*, 41(2) :93–137, 1987.

SESSION 2 : Apprentissage automatique et regroupement présidée par Florence Dupin de Saint-Cyr

Des proportions logiques à un nouvel algorithme de clustering

Myriam Bounhas[1], Henri Prade[2]

[1] LARODEC, ISG, Université de Tunis, Tunisie
[2] IRIT, Univ. Toulouse, Toulouse INP, CNRS, Toulouse, France

myriam_bounhas@yahoo.fr, henri.prade@irit.fr

Résumé

Cet article propose une nouvelle méthode de regroupement (clustering) capable de rivaliser avec des algorithmes comme HDBSCAN ou k-means. L'algorithme regroupe des paires d'éléments plutôt que des éléments isolés et s'appuie sur deux proportions logiques : la paralogie et la paralogie inverse. Ce sont des connecteurs logiques booléens, quaternaires, qui peuvent être étendus à des valeurs numériques. La paralogie évalue dans quelle mesure deux paires d'éléments partagent des caractéristiques similaires, suggérant qu'elles peuvent appartenir au même cluster. En revanche, la paralogie inverse évalue si ces paires sont orthogonales et devraient appartenir à des clusters différents. L'algorithme commence par créer de petits clusters à l'aide d'un indice de paralogie inverse, puis les fusionne progressivement grâce à un indice de paralogie.

Mots-clés

Regroupement, Paralogie, Inverse Paralogie.

Abstract

This article proposes a new clustering method capable of competing with algorithms such as HDBSCAN or k-means. The algorithm clusters pairs of elements rather than isolated elements and is based on two logical proportions : paralogy and inverse paralogy. These are Boolean, quaternary logical connectors that can be extended to numerical values. Paralogy assesses the extent to which two pairs of elements share similar characteristics, suggesting that they may belong to the same cluster. In contrast, inverse paralogy assesses whether these pairs are orthogonal and should belong to different clusters. The algorithm begins by creating small clusters using a reverse paralogy index, then gradually merges them using a paralogy index.

Keywords

Clustering, Paralogy, Inverse Paralogy.

1 Introduction

La formalisation logique du concept de proportions analogiques, qui sont des énoncés de la forme « a est à b comme c est à d », a conduit à la définition des proportions logiques [8]. Les proportions logiques comparent la similitude / dissemblance des éléments d'une paire avec la similitude / dissemblance des éléments d'une autre paire. Parmi les exemples de telles proportions, on peut citer les proportions analogiques qui expriment que « a diffère de b comme c diffère de d et b diffère de a comme d diffère de c », ainsi que d'autres proportions remarquables.

Parmi celles-ci, celle appelée « paralogie » présente un intérêt particulier : elle exprime que « a et b sont similaires comme c et d sont similaires », où la similitude peut faire référence à des caractéristiques présentes ou absentes, établissant ainsi un parallèle entre les paires (a, b) et (c, d). Une autre, appelée « paralogie inverse » [9], exprime plutôt une forme d'orthogonalité entre les paires : « ce que a et b ont en commun, c et d ne l'ont pas, et vice versa ».

Ces deux proportions logiques nous rappellent ce qui caractérise l'idée de cluster. En effet, en apprentissage non supervisé, les techniques de regroupement visent à structurer un ensemble de données en groupes de telle sorte que les éléments d'un groupe aient plus en commun entre eux qu'avec les autres éléments, c'est-à-dire qu'ils soient plus similaires entre eux qu'avec ceux d'autres groupes. Cela suggère la possibilité de construire un algorithme de regroupement qui repose sur l'utilisation de ces proportions. C'est ce que nous explorons dans cet article.

Le document est organisé comme suit. Après une brève présentation des proportions logiques en Section 2, la notion plus ciblée de comparateur de paires est introduite en Section 3 et appliquée à la paralogie et à la paralogie inverse, d'abord dans un cadre booléen, puis étendue en logique multivalente. La Section 4 présente un nouvel algorithme de regroupement basé sur la paralogie [1], et la Section 5 rend compte d'expériences ayant donné des résultats compétitifs. La Section 6 compare les résultats obtenus par l'approche proposée à ceux obtenus avec HDBSCAN et la méthode $k\text{-}means$.

2 Proportions logiques : une brève présentation

Les proportions logiques booléennes [9, 11] sont des connecteurs T reliant quatre variables booléennes, disons a, b, c, d, qui lient par une équivalence deux indicateurs de comparaison entre a et b à deux indicateurs de comparaison entre c et d. Les indicateurs de comparaison sont notés

1. Un premier algorithme exclusivement basé sur la paralogie inverse a été présenté dans [2] et ses performances y sont comparées à l'approche $k\text{-}means$ uniquement.

ci dans l'expression (*) ci-après. Plus précisément, une proportion logique T est de la forme

$$T(a,b,c,d) = [ci_1(a,b) \equiv ci_2(c,d)] \wedge [ci_3(a,b) \equiv ci_4(c,d)] \quad (*)$$

où chacun des ci_i est l'un des quatre indicateurs de comparaison booléenne possibles, à savoir $ci_i(a,b)$ est l'un quelconque parmi $a \wedge b$, $\neg a \wedge \neg b$, $\neg a \wedge b$ ou $a \wedge \neg b$ (avec $ci_1 \neq ci_3$ ou $ci_2 \neq ci_4$ pour s'assurer que la conjonction n'est pas triviale). Notez que les deux premiers indicateurs, $a \wedge b$ et $\neg a \wedge \neg b$, évaluent respectivement la similarité positive et négative, tandis que les deux derniers $\neg a \wedge b$ et $a \wedge \neg b$ sont des indicateurs de dissemblance. Prenons les exemples de la paralogie et de la paralogie inverse :

$$P(a,b,c,d) = [(a \wedge b) \equiv (c \wedge d)] \wedge [(\neg a \wedge \neg b) \equiv (\neg c \wedge \neg d)]$$

$$I(a,b,c,d) = [(a \wedge b) \equiv (\neg c \wedge \neg d)] \wedge [(\neg a \wedge \neg b) \equiv (c \wedge d)]$$

La paralogie P stipule clairement que « ce que a et b ont en commun, c et d l'ont également, positivement ou négativement ». La paralogie inverse I affirme le contraire : « lorsque a et b sont vrais, c et d sont faux, et vice-versa ». Comme toute proportion logique, P et I ne sont vraies que pour 6 valuations (parmi les 2^4 possibles) [10], données dans la Table 1.

TABLE 1 –Tables de vérité : Paralogie (**P**), Paralogie Inverse (**I**)

				P			
a	b	c	d	a	b	c	d
0	0	0	0	0	0	1	1
1	1	1	1	1	1	0	0
0	1	1	0	0	1	1	0
1	0	0	1	1	0	0	1
0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0

Il convient de noter que la paralogie est étroitement liée à la proportion analogique A par une simple permutation, puisque nous avons $A(a,b,c,d) \equiv P(a,d,c,b)$ [11].

3 Comparateurs de paires

Les paires peuvent être comparées non seulement à l'aide des équivalences $\equiv$ comme dans (*), mais aussi à l'aide du connecteur opposé, l'opérateur du « ou exclusif » $\oplus$. Ainsi, si nous substituons $\oplus$ à $\equiv$ dans (*), T devient

$$\mathcal{S}^{\oplus}(T)(a,b,c,d) = [ci_1(a,b) \oplus ci_2(c,d)] \wedge [ci_3(a,b) \oplus ci_4(c,d)] \quad (**)$$

Pour une étude systématique des comparateurs de paires, obtenus par cette substitution à partir des proportions logiques, le lecteur pourra consulter [12].
Si nous appliquons la substitution $\mathcal{S}^{\oplus}$ à P et I, nous obtenons :
$\underline{\underline{P}}(a,b,c,d) = [(a \wedge b) \oplus (\neg c \wedge \neg d)] \wedge [(\neg a \wedge \neg b) \oplus (c \wedge d)]$
$\underline{\underline{I}}(a,b,c,d) = [(a \wedge b) \oplus (c \wedge d)] \wedge [(\neg a \wedge \neg b) \oplus (\neg c \wedge \neg d)] =$
$[(a \wedge b) \oplus (c \wedge d)] \wedge [(a \vee b) \oplus (c \vee d)]$
où
$\underline{\underline{P}}(a,b,c,d) = \mathcal{S}^{\oplus}(I)(a,b,c,d)$,
$\underline{\underline{I}}(a,b,c,d) = \mathcal{S}^{\oplus}(P)(a,b,c,d)$.
Les tables de vérité de $\underline{\underline{P}}$ et $\underline{\underline{I}}$ sont données dans la Table 2, où seuls les deux valuations qui les rendent vrais apparaissent, tandis que les 14 autres valuations possibles rendent ces connecteurs faux.

TABLE 2 – Tables de vérité de $\underline{\underline{P}}$ et $\underline{\underline{I}}$

				$\underline{\underline{P}}$			
a	b	c	d	a	b	c	d
0	0	0	0	0	0	1	1
1	1	1	1	1	1	0	0

Comme on peut le voir, $\underline{\underline{P}}$ et $\underline{\underline{I}}$ sont beaucoup plus ciblés que P et I, puisqu'ils ne conservent que les deux premières lignes de la Table 1, omettant les évaluations communes à P et I et ne conservant que les valuations qui expriment leur véritable identité.
Les expressions booléennes de P et I peuvent être étendues à des connecteurs à valeurs multiples avec des valeurs dans l'intervalle $[0,1]$. Nous utilisons les connecteurs de Łukasiewicz [5], à savoir

- $\neg x = 1 - x$;
- $x \wedge y = \min(x,y)$; $x \vee y = \max(x,y)$;
- $x \rightarrow y = \min(1, 1 - x + y)$;
- $x \equiv y = (x \rightarrow y) \wedge (y \rightarrow x) = 1 - |x - y|$;
- $x \oplus y = \neg(x \equiv y) = |x - y|$.

Cela conduit aux extensions suivantes pour $\underline{\underline{P}}$ et $\underline{\underline{I}}$, en conservant *la même structure de comparaison par paires* :
Pour $(x,y,z,t) \in [0,1]^4$,

$$\underline{\underline{P}}(x,y,z,t) = \min(|\min(x,y) - \min(1-z, 1-t)|, |\min(1-x, 1-y) - \min(z,t)|) \quad (1)$$

De façon similaire, on obtient pour $\underline{\underline{I}}$

$$\underline{\underline{I}}(x,y,z,t) = \min(|\min(x,y) - \min(z,t)|, |\max(x,y) - \max(z,t)|) \quad (2)$$

Il existe une autre façon d'écrire ces deux connecteurs qui ne sont vrais que pour les deux valuations de la Table 2 :

$$\underline{\underline{P}}(a,b,c,d) = (a \equiv b) \wedge (c \equiv d) \wedge (a \equiv c) \wedge (b \equiv d)$$

$$\underline{\underline{I}}(a,b,c,d) = (a \equiv b) \wedge (c \equiv d) \wedge (a \equiv \neg c) \wedge (b \equiv \neg d)$$

La structure des formules ci-dessus diffère clairement de celle obtenue à partir de (**), bien qu'elles aient les mêmes tables de vérité : les expressions ne se concentrent pas uniquement sur les paires (a,b) et (c,d), mais impliquent également les paires (a,c) et (b,d). L'expression pour $\underline{\underline{P}}$ s'inscrit dans l'esprit de la vision en termes de parallélogramme des proportions analogiques [13] (considérant les paires comme des côtés). La condition $b \equiv d$ est superflue ($\equiv$

est transitif), mais a été ajoutée pour un traitement égal des littéraux. Cela conduit aux extensions graduelles suivantes :

$$\underline{\underline{P}}(x,y,z,t) = 1-\max(|x-y|,|z-t|,|x-z|,|y-t|) \quad (3)$$

$$\underline{\underline{I}}(x,y,z,t) = 1-\max(|x-y|,|z-t|,|x-1+z|,|y-1+t|) \quad (4)$$

Notons que (2) et (3) n'impliquent pas de négation interne, contrairement à (1) et (4) : nous privilégions donc leur utilisation dans la pratique. $\underline{\underline{P}}$ diminue dès qu'une valeur s'écarte des autres.

La paralogie et l'orthogonalité s'étendent aux *paires de vecteurs*, avec n composantes valuées dans $[0,1]$:

$$\mathcal{P}((\vec{x},\vec{y}),(\vec{z},\vec{t})) = \sum_{i=1}^{n} \underline{\underline{P}}(x_i,y_i,z_i,t_i)/n \quad (5)$$

et

$$\mathcal{O}((\vec{x},\vec{y}),(\vec{z},\vec{t})) = \sum_{i=1}^{n} \underline{\underline{I}}(x_i,y_i,z_i,t_i)/n. \quad (6)$$

4 Algorithme de regroupement basé sur la paralogie

Nous présentons une nouvelle procédure de regroupement qui opère sur des paires de points plutôt que sur des points individuels. Elle utilise la paralogie $\mathcal{P}$ entre les paires (formule (5)) et l'orthogonalité $\mathcal{O}$ entre les paires (formule (6)) , définies à la fin de la Section 3. $\mathcal{P}$ estime la cohérence structurelle entre deux paires, et $\mathcal{O}$ leur séparation.

Notations et préliminaires. Soit $\mathcal{X} = \{x_1,\ldots,x_N\} \subset \mathbb{R}^d$ l'ensemble de données. Soit $\mathbb{P} = \{(a^j,b^j)\}_{j=1}^m$ un ensemble de paires de points (représentant des données), généré dans une étape préliminaire. Chaque paire $p = (a^j,b^j) \in \mathbb{P}$ est définie de telle sorte que b^j soit le plus proche voisin de a^j par rapport à la distance euclidienne, et que les doublons ou chevauchements de paires soient évités. À une itération donnée, le regroupement est représenté par une partition $\mathcal{C} = \{C_1,\ldots,C_k\}$.

4.1 Procédure de base

L'algorithme fonctionne en créant, ajustant et consolidant, de manière répétée, des (petits) clusters intermédiaires. Ce processus implique : (i) la détection de paires de points qui présentent une forte orthogonalité, (ii) placer ces paires dans des mini-clusters initiaux et provisoires, et (iii) fusionner progressivement les mini-clusters dont les contenus sont mutuellement cohérents en termes de paralogie, ce qui garantit "la cohérence relationnelle" en termes de "similarité".

La consolidation des clusters obéit à une règle de préservation de la cohérence régie par un paramètre de tolérance α spécifié par l'utilisateur, garantissant que seuls les mini-clusters mutuellement compatibles sont fusionnés. Nous fournissons maintenant une description détaillée de chaque étape.

Prétraitement des données. Soit $\mathcal{X}$ un ensemble de données de points dans un espace à N dimensions. Dans un premier temps, toutes les caractéristiques sont mises à l'échelle afin que leurs valeurs se situent dans des plages comparables. Cette normalisation est essentielle pour garantir une utilisation homogène des indices de paralogie et d'orthogonalité entre les dimensions. Ensuite, l'ensemble $\mathbb{P}$ des paires de points les plus proches est généré.

Calcul des matrices d'orthogonalité et de paralogie. Au cours d'une phase préliminaire hors ligne, l'indice d'orthogonalité est évalué pour chaque paire ordonnée de paires $(p^j,q^j) \in \mathbb{P}\times\mathbb{P}$ à l'aide de la définition donnée à la fin de la section 3. Les valeurs obtenues sont stockées dans la matrice d'orthogonalité $\boldsymbol{\mathcal{O}}$. Une procédure analogue est effectuée pour calculer et stocker les indices de paralogie correspondants dans la matrice de paralogie $\boldsymbol{\mathcal{P}}$.

Initialisation des mini-clusters. Cette étape commence par la sélection des deux paires qui présentent le plus haut degré d'orthogonalité selon la matrice $\boldsymbol{\mathcal{O}}$. Ces paires définissent les deux premiers mini-clusters et leurs centroïdes initiaux, représentant la configuration la plus dissemblable sur le plan structurel. D'autres mini-clusters sont introduits progressivement en identifiant les paires qui sont suffisamment orthogonales par rapport à tous les centroïdes existants. Une paire $p \in \mathbb{P}$ est choisie comme nouveau centroïde lorsque son orthogonalité par rapport à chaque centroïde actuel dépasse un seuil donné β (β est maintenu faible, par exemple 0,1).

Affectation des paires aux clusters. Chaque paire restante non affectée $p \in \mathbb{P}$ est successivement attribuée au mini-cluster dont le centroïde donne la plus petite valeur d'orthogonalité avec p. Le recours à des comparaisons basées sur les centroïdes réduit considérablement la charge de calcul en évitant les comparaisons exhaustives par paires au sein des clusters.

Étape de raffinement. Régulièrement, les centroïdes des mini-clusters sont recalculés en faisant la moyenne des coordonnées des points qui constituent les paires attribuées à chaque cluster. Cette mise à jour est effectuée après qu'une fraction spécifiée ($level\ l$) du total des paires a été traitée (par exemple, $l = 50\%$ ou 75%).

Fusion des mini-clusters. Une fois que toutes les paires ont été attribuées à des mini-clusters, une phase de fusion est appliquée afin de réduire la sur-segmentation et de produire des clusters plus stables. Cette phase suit une stratégie agglomérative contrainte par un critère de préservation de la cohérence. Soit $\mathcal{C} = \{C_1,\ldots,C_k\}$ l'ensemble actuel de mini-clusters et soit $\mathbb{P}_i \subset \mathbb{P}$ le sous-ensemble de paires dont les deux points appartiennent au cluster C_i.

Cohérence intra-cluster. Pour chaque mini-cluster C_i, un score de cohérence intra-cluster est défini comme la paralogie moyenne entre toutes les paires distinctes de l'ensenble

$\mathbb{P}_i$ appartenant au cluster C_i :

$$\mathrm{coh}_{\mathrm{intra}}(C_i) = \frac{1}{|\mathbb{P}_i|(|\mathbb{P}_i|-1)} \sum_{\substack{p,q \in \mathbb{P}_i \\ p \neq q}} \mathcal{P}(p,q), \qquad (7)$$

où $\mathcal{P}$ est la mesure de paralogie entre deux paires. Ce score quantifie l'homogénéité structurelle interne de C_i.

Étant donnés deux mini-clusters distincts C_i et C_j, leur similitude est approximée par la paralogie entre leurs centroïdes respectifs :

$$\mathrm{coh}_{\mathrm{inter}}(C_i, C_j) \approx \mathcal{P}(\mu_i, \mu_j), \qquad (8)$$

où μ_i et μ_j désignent les centroïdes (paires de points) associés aux clusters C_i et C_j. Cette approximation basée sur les centroïdes évite le coût computationnel lié à l'évaluation de la paralogie entre toutes les paires de tous les clusters.

Critère de fusion. Une fusion entre les clusters C_i et C_j est considérée comme admissible si elle ne détériore pas la cohérence interne de chaque cluster individuel, c'est-à-dire si la similarité inter-clusters dépasse la cohérence interne maximale entre les deux clusters individuels (avec un certain degré de tolérance α) :

$$\mathrm{coh}_{\mathrm{inter}}(C_i, C_j) > \alpha * \max\left(\mathrm{coh}_{\mathrm{intra}}(C_i), \mathrm{coh}_{\mathrm{intra}}(C_j)\right) \qquad (9)$$

où $\alpha > 0$ est un paramètre de tolérance défini par l'utilisateur qui contrôle la rigueur de la préservation de la cohérence.

Parmi toutes les paires de clusters admissibles, on choisit la fusion qui maximise $\mathrm{coh}_{\mathrm{inter}}$. Après la fusion, les étiquettes des clusters sont mises à jour, le centroïde du cluster fusionné est recalculé, ainsi que sa cohérence intra-cluster. La procédure est répétée de manière itérative jusqu'à ce qu'il n'y ait plus de fusion admissible.

Arrêt. L'algorithme se termine lorsqu'aucune paire de mini-clusters ne satisfait au critère de fusion préservant la cohérence. Le résultat final comprend le nombre final de clusters $k^\star$, les centroïdes finaux des clusters $\mu^\star$ et les affectations finales des clusters $\mathcal{C}^\star$ pour tous les points .

4.2 Algorithme

L'algorithme 1 décrit la procédure principale de l'approche proposée. Cet algorithme appelle la fonction $init_centroids$ pour initialiser les mini-clusters et la fonction $merge_clusters$ pour les fusionner lorsque cela est nécessaire, comme expliqué ci-dessus.

Algorithm 1 Paralogy, un Algorithme de Clustering

Input : $\mathcal{X}$: Données, β : seuil init. , α : seuil fusion.
Output : $k^\star, \mathcal{C}^\star, \mu^\star$: nb final de clusters, clusters, et centroïdes.

$\mathcal{X} \leftarrow$ Normalize($\mathcal{X}$) {Prétraitement}
$\mathbb{P} \leftarrow \{(a,b) \mid a,b \in \mathcal{X} \times \mathcal{X}, b =$ Nearest-Neighbor$(a)\}$
$\boldsymbol{\mathcal{O}} \leftarrow \{\mathcal{O}_{i,j} \mid p_i, p_j \in \mathbb{P}^2, \mathcal{O}_{i,j} = \mathcal{O}((a,b),(c,d))\}$
$\boldsymbol{\mathcal{P}} \leftarrow \{\mathcal{P}_{i,j} \mid p_i, p_j \in \mathbb{P}^2, \mathcal{P}_{i,j} = \mathcal{P}((a,b),(c,d))\}$ {Matrices}
$k, \mathcal{C}, \mu \leftarrow$ init_centroids($\boldsymbol{\mathcal{O}}$) {Initialisation}
for each unclustered $p = (a,b) \in \mathbb{P}$ **do** {Assignation}
  $C \leftarrow \arg\min_j(\mathcal{O}((a,b),(c,d))|\mu_j = (c,d),\ j = 1,\ldots,k)$; C.append(p)
end for
if $|$Clustered$(\mathbb{P})|/|\mathbb{P}| > l$ **then** UpdateCentroids(μ) {Refine}
$k^\star, \mathcal{C}^\star, \mu^\star \leftarrow$ merge_clusters$(k, \mathcal{C}, \mu, \alpha)$ {Fusion}
return $k^\star$, $\mathcal{C}^\star$, $\mu^\star$

5 Résultats expérimentaux

Cette Section présente une analyse des résultats empiriques obtenus à l'aide de la méthode proposée de *regroupement basé sur la paralogie* sur des ensembles de données synthétiques ou réels. Nous comparons également l'efficacité de cette méthode avec l'algorithme HDBSCAN [4, 3].

Nous expérimentons une variété d'ensembles de données présentant des caractéristiques différentes : un ensemble de données synthétiques (Blobs) et trois ensembles de données réelles (Iris, Zoo et Breast Cancer).

Pour évaluer les deux méthodes de regroupement comparées, nous utilisons trois mesures d'évaluation : le *score Silhouette (Sil)*, l'*indice de Calinski-Harabasz (CH)* et le *score de Davies-Bouldin (DB)* pour les ensembles de données d'apprentissage non supervisé (Blobs), et nous appliquons l'*Adjusted Rand Index (ARI)* plus adapté aux trois ensembles de données d'apprentissage supervisé. Des valeurs plus élevées (proches de 1 pour Sil et ARI) pour Sil, CHI et ARI sont préférables, tandis que des valeurs plus faibles (proches de 0) pour DB sont préférables. Les résultats sont présentés respectivement dans en Figure 1 pour Blobs, en Figure 2 pour Iris, en Figure 3 pour Zoo et en Figure 4 pour les ensembles de données sur le cancer du sein. Ces résultats correspondent aux meilleurs paramètres obtenus pour chaque algorithme. Pour le regroupement basé sur la paralogie, nous affichons les centroïdes des mini-clusters (sous forme d'étoiles noires) afin de suivre le processus de fusion. Nous indiquons d'abord les performances de l'algorithme proposé pour chaque ensemble de données.

5.1 Blobs - 5 amas gaussiens isotropes

Dans les ensembles de données Blobs, tout point d'un cluster est plus proche de son centroïde que du centroïde de tout autre cluster. L'algorithme Paralogy obtient des mesures de validation meilleures (Sil = 0,80, CHI = 10433,6,

DB = 0,27) que HDBSCAN (Sil= 0,76, CHI = 8404,6, DB = 0,98), ce qui indique un meilleur regroupement. Alors que HDBSCAN reste sensible aux variations de densité locales et introduit un certain bruit (2 points noirs), l'algorithme Paralogy préserve l'intégrité des clusters en s'appuyant sur la cohérence relationnelle plutôt que sur les écarts de densité. Les centroïdes s'alignent correctement avec la structure réelle des blobs, démontrant ainsi sa robustesse pour cet ensemble de données.

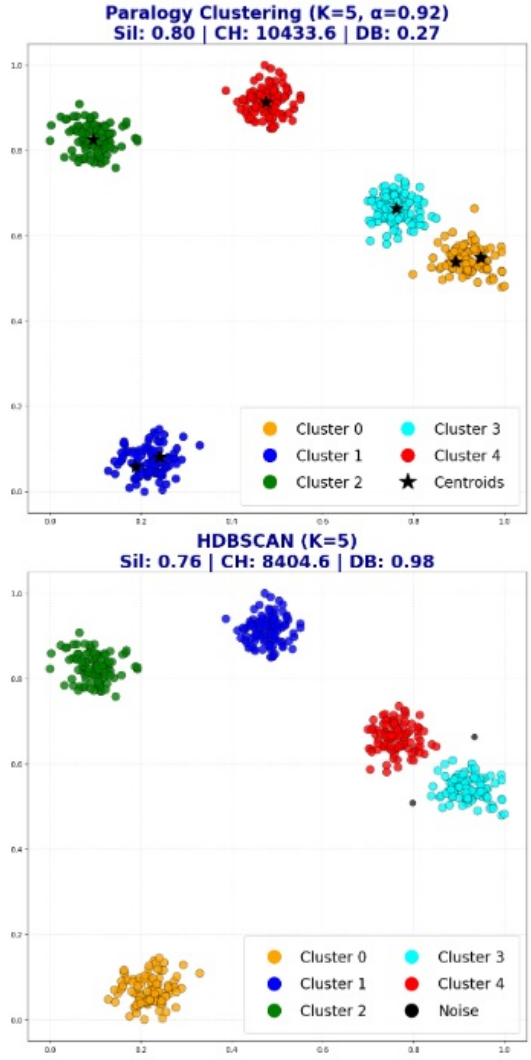


FIGURE 1 – Résultats pour l'ensemble de données Blobs

5.2 Iris - Relations d'échelle allométriques

L'ensemble de données Iris contient 3 classes, dont l'une est linéairement séparable des deux autres ; ces dernières ne sont pas linéairement séparables l'une de l'autre. L'algorithme Paralogy obtient un score ARI beaucoup plus élevé (0,886) que HDBSCAN (0,568). Bien que les espèces Versicolor et Virginica se chevauchent fortement dans l'espace, chaque espèce suit une relation de croissance allométrique distincte entre la longueur et la largeur des pétales. Même lorsque les échantillons sont proches en distance absolue, leurs taux de croissance locaux diffèrent. L'algorithme Paralogy capture ces caractéristiques invariables et réussit à séparer les espèces, tandis que HDBSCAN traite le chevauchement comme une seule région à haute densité, empêchant ainsi la séparation. pour cet ensemble de données.

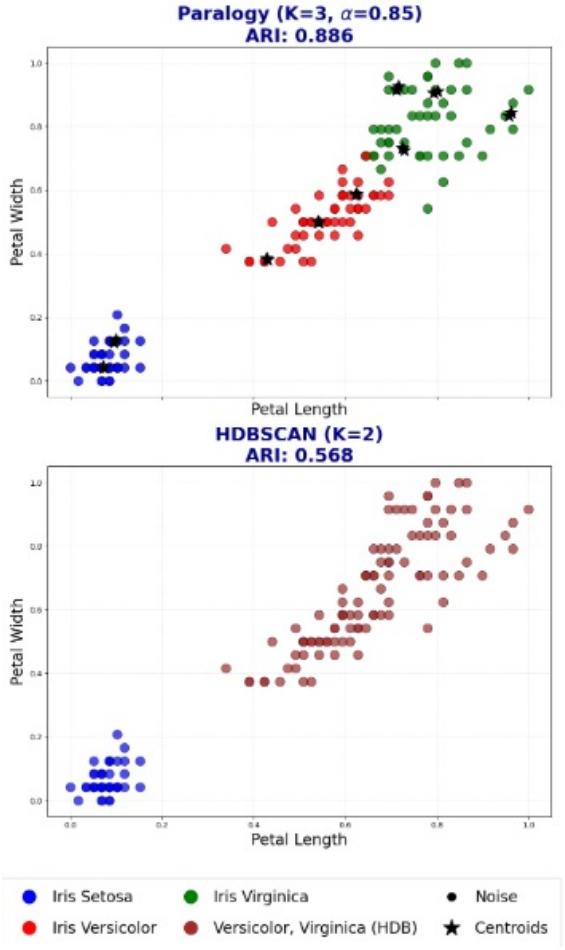


FIGURE 2 – Résultats pour l'ensemble de données Iris

5.3 Zoo - Espace discret avec descriptions booléennes

Dans cet ensemble de données, 101 animaux appartenant à 7 classes sont décrits à l'aide de 16 attributs booléens. L'algorithme Paralogy surpasse HDBSCAN avec un ARI de 0,868 contre 0,475 (les écarts de densité sont mal définis). Paralogy maintient la cohérence relationnelle au sein des classes biologiques. Cette approche préserve correctement le groupe des mammifères, même pour des anomalies comme l'ornithorynque, qui est regroupé correctement grâce à des traits décisifs (poils, lait) malgré ses caractéristiques reptiliennes (voir le point brun au centre). De plus, Paralogy réussit à démêler les groupes qui se chevauchent comme les amphibiens et les reptiles, en identifiant huit clusters cohérents. En revanche, HDBSCAN a du mal dans cet espace discret ; il fragmente la classe des mammifères, crée des clusters supplémentaires arbitraires, fusionne inutilement les amphibiens et les reptiles et étiquette à tort comme du bruit des échantillons isolés mais logiquement cohérents.

5.4 Breast Cancer Wisconsin

Dans cet ensemble de données à deux classes (malignes vs. bénignes) avec de multiples dimensions (30 caractéristiques) présentant un chevauchement clair des clusters, Paralogy (ARI = 0,501) surpasse largement HDBSCAN (ARI = 0,168). Au lieu d'effectuer un regroupement basé sur la densité, Paralogy capture les lois de corrélation stables entre les caractéristiques des noyaux cellulaires (par exemple, le rayon et la surface), ce qui nous permet de séparer les échantillons qui se chevauchent spatialement mais qui diffèrent relationnellement. Il maintient la continuité des classes malgré la présence de bruit et identifie un petit sous-groupe cohérent supplémentaire (Extra Cluster 1) qui peut représenter un sous-groupe spécial cohérent. En revanche, HDBSCAN fragmente la structure et étiquette de nombreux points ambigus comme du bruit.

6 Discussion / Comparaison avec HDBSCAN et k-means

Contrairement à HDBSCAN, qui dépend de la densité des points et peut échouer lorsque les écarts de densité entre les groupes sont faibles ou inexistants (comme dans le cas d'Iris ou de Breast Cancer), Paralogy privilégie la cohérence relationnelle plutôt que la proximité spatiale. En passant d'une approche basée sur la proximité à un paradigme structurel relationnel, il construit des clusters en fonction de caractéristiques communes. Cela lui permet de préserver l'intégrité des groupes indépendamment de la densité locale, en intégrant les points isolés ou en séparant les groupes qui se chevauchent, à condition qu'ils suivent une logique interne cohérente. En outre, en cas d'ambiguïté, HDBSCAN peut laisser de nombreux points en dehors des clusters en tant que bruit, ce qui n'est pas le cas de la méthode basée sur la paralogie qui affecte tous les points à des clusters.

Contrairement à k-$means$ [7, 6], qui nécessite de spécifier à l'avance le nombre de clusters k (par exemple, via la méthode "Elbow"), Paralogy estime le nombre de clusters directement pendant le processus de clustering, sans aucune procédure supplémentaire. Les résultats préliminaires indiquent que Paralogy surpasse nettement k-$means$ sur les ensembles de données Blobs et Iris (pour Blobs, k-$means$ ne montre que trois clusters au lieu de cinq avec des scores Sil = 0,74, CH = 1925,4, DB = 0,43 et pour Iris seulement 2 clusters avec ARI= 0,568) et reste compétitif sur le jeu de données Moons, où les méthodes basées sur la densité capturent mieux les deux demi-lunes.

Pour traiter les structures non convexes (par exemple Moons), nous avons développé une variante de l'algorithme Paralogy. Au lieu d'estimer la cohérence inter-clusters uniquement à partir des scores de paralogie des centroïdes, nous (i) définissons 2 ou 3 paires de limites au sein de chaque mini-cluster et (ii) calculons la cohérence inter-clusters comme le score de paralogie maximal sur les paires de limites. Le nouveau critère de fusion garantit que seuls les mini-clusters présentant une forte cohérence des limites sont fusionnés. Les résultats montrent que cette stra-

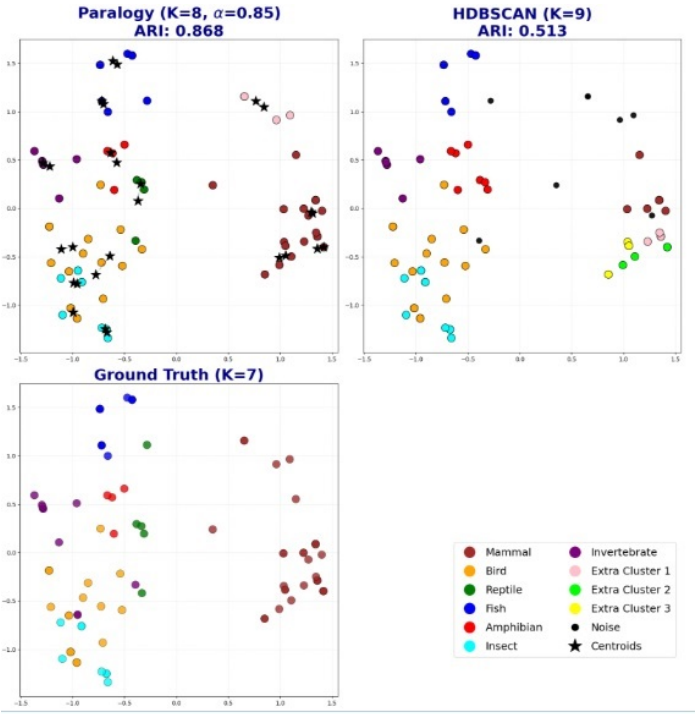


FIGURE 3 – Résultats pour l'ensemble de données Zoo

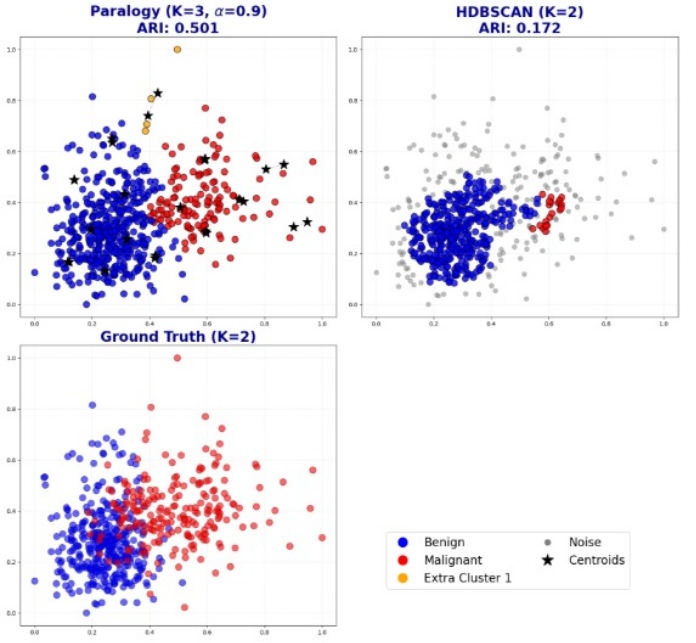


FIGURE 4 – Résultats pour l'ensemble de données Breast Cancer

tégie identifie les deux structures en demi-lune dans l'ensemble de données Moons. L'évaluation de son efficacité sur d'autres ensembles de données fera l'objet de travaux futurs.
Les expérimentations à l'appui de la discussion ci-dessus sont données en Section 8.

7 Conclusion

Une analyse logique de l'idée de cluster a conduit à une nouvelle méthode de regroupement. Le regroupement basé sur la paralogie introduit une nouvelle perspective en apprentissage non supervisé en identifiant les clusters grâce à la cohérence relationnelle plutôt qu'en se basant uniquement sur la densité ou la distance.
Dans l'ensemble, les résultats expérimentaux suggèrent que cette approche est particulièrement intéressante pour les données dont l'organisation est régie par des relations sémantiques, permettant l'émergence de clusters significatifs même en l'absence de séparations claires en termes de densité. Les méthodes basées sur la densité, telles que HDBSCAN, restent compétitives pour les structures essentiellement définies par la densité. Plutôt que de se faire concurrence directement, les deux approches doivent être considérées comme complémentaires : l'une découvre les clusters comme des régions denses, l'autre comme des groupes régis par des relations sémantiques locales cohérentes.
Il serait aussi intéressant d'examiner si les idées de « clusters flous », à la base de l'algorithme « fuzzy c-means » [1], peuvent être introduites avec profit dans l'approche présentée ici.

8 Annexe : Résultats supplémentaires concernant la discussion de la Section 6

Dans cet annexe, nous fournissons les figures supplémentaires qui appuient les affirmations faites dans la Section discussion : i) concernant k-means pour les jeux de données Blobs et Iris et ii) concernant la variante de l'algorithme Paralogy exploitant des paires de frontière sur Moons.

8.1 Comparaison avec k-means

Les résultats de l'étude comparative avec l'algorithme k-means sont présentés dans la Figure 5 pour Blobs et la Figure 6 pour Iris. Nous rappelons d'abord que k-means exploite la méthode du coude (Elbow) pour estimer le nombre de clusters. Comme indiqué précédemment dans la Section 5, la Paralogy montre une forte capacité à détecter le nombre correct de clusters pour le jeu de données Blobs. En particulier, l'approche proposée surpasse k-means pour toutes les métriques utilisées. En revanche, la méthode du Elbow sous-estime le nombre de clusters, ce qui force par conséquent k-means à fusionner « inutilement » certains clusters distincts.
De plus, comme illustré dans la Figure 6, la Paralogy présente une capacité plus élevée à retrouver la structure intrinsèque du jeu de données Iris ($k = 3$, ARI = 0.886) comparé à k-means lorsque ce dernier est contraint par le critère de sélection du Elbow. Bien que l'heuristique du Elbow identifie $k = 2$ comme optimal, ce choix conduit à une sous-estimation claire du nombre réel de clusters, forçant à regrouper les espèces Versicolor et Virginica dans une seule partition (ARI = 0.568).

8.2 Paralogy : Stratégies Centroïde vs Frontière

Pour traiter des structures non convexes, dans l'expérience Moons (Figure 7), nous avons développé une variante de l'algorithme de paralogie exploitant des paires de frontière. Au lieu d'utiliser uniquement les scores de paralogie des centroïdes, nous mesurons la cohérence inter-clusters par le score maximal entre les paires de frontière de chaque mini-cluster. Dans la Figure 7, les points jaunes indiquent les paires de frontière utilisées lors de la dernière fusion.
Les deux stratégies — basée sur le centroïde et sur les paires de frontière — cherchent à capturer la séparation entre les deux arcs non convexes, mais présentent des sensibilités différentes à la structure des données. La stratégie par centroïde fournit une représentation globale stable, tandis que la stratégie de frontière reflète mieux la configuration structurelle locale. Empiriquement, cette approche montre des résultats prometteurs pour identifier les deux arcs, suggérant que l'intégration d'informations de frontière peut améliorer le clustering de structures non convexes.

Remerciements

Cette recherche a été financée par le projet de l'ANR "Analogies: from theory to tools and applications" (AT2TA), ANR-22-CE23-0023.

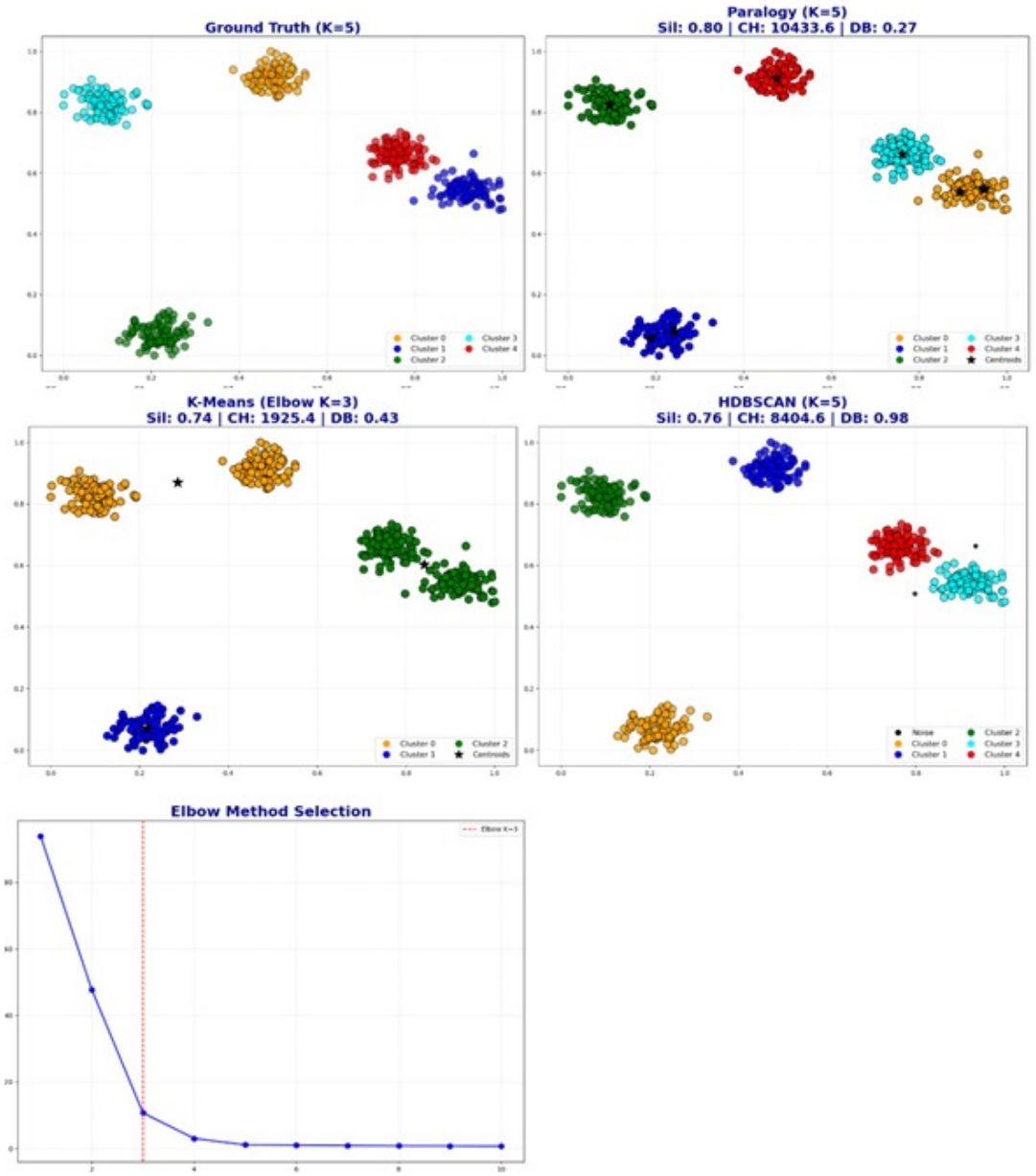


FIGURE 5 – Paralogy/HDBSCAN/k-means : Résultats pour le jeu de données Blobs

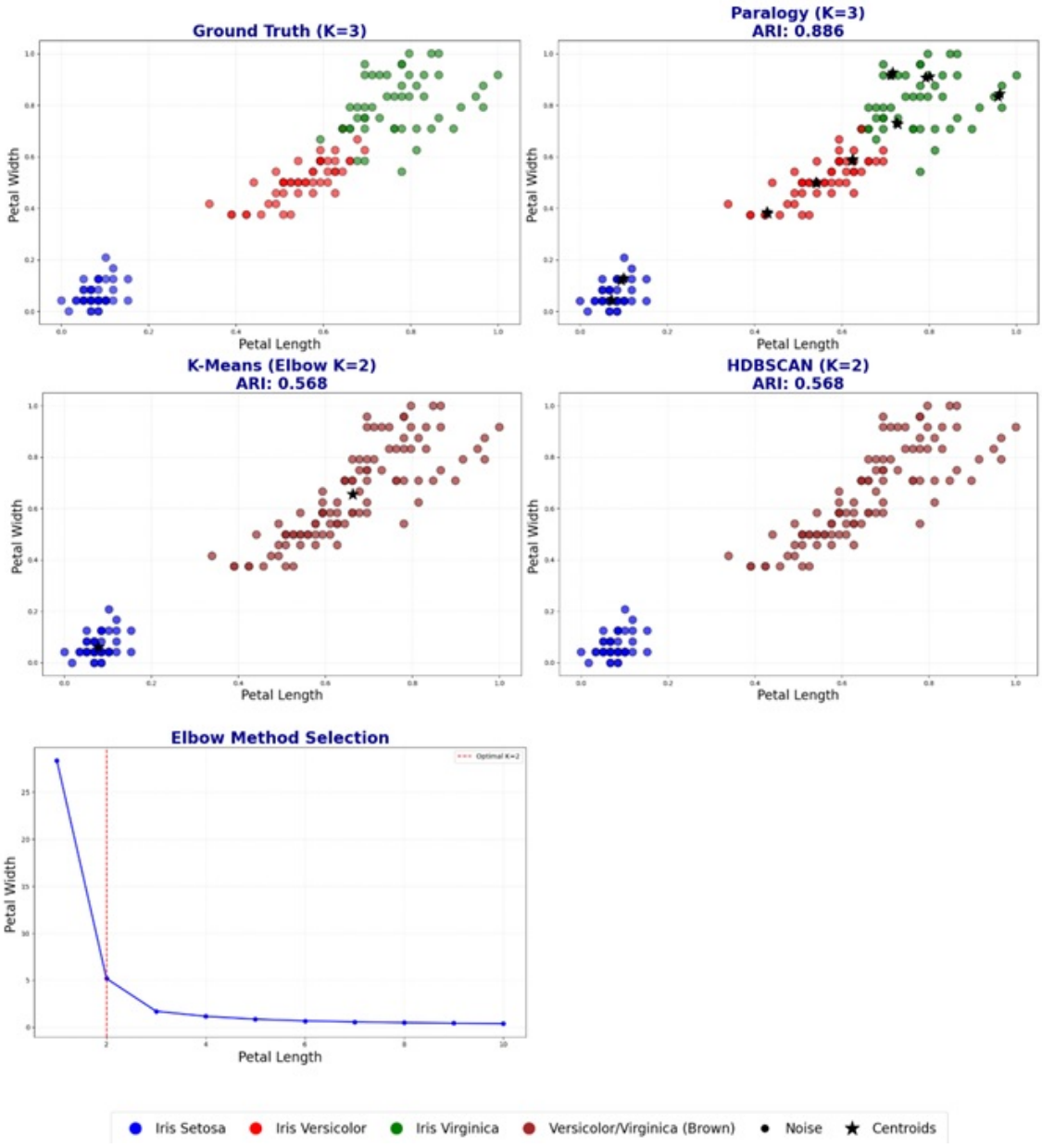


FIGURE 6 – Paralogy/HDBSCAN/k-means : Résultats pour le jeu de données Iris

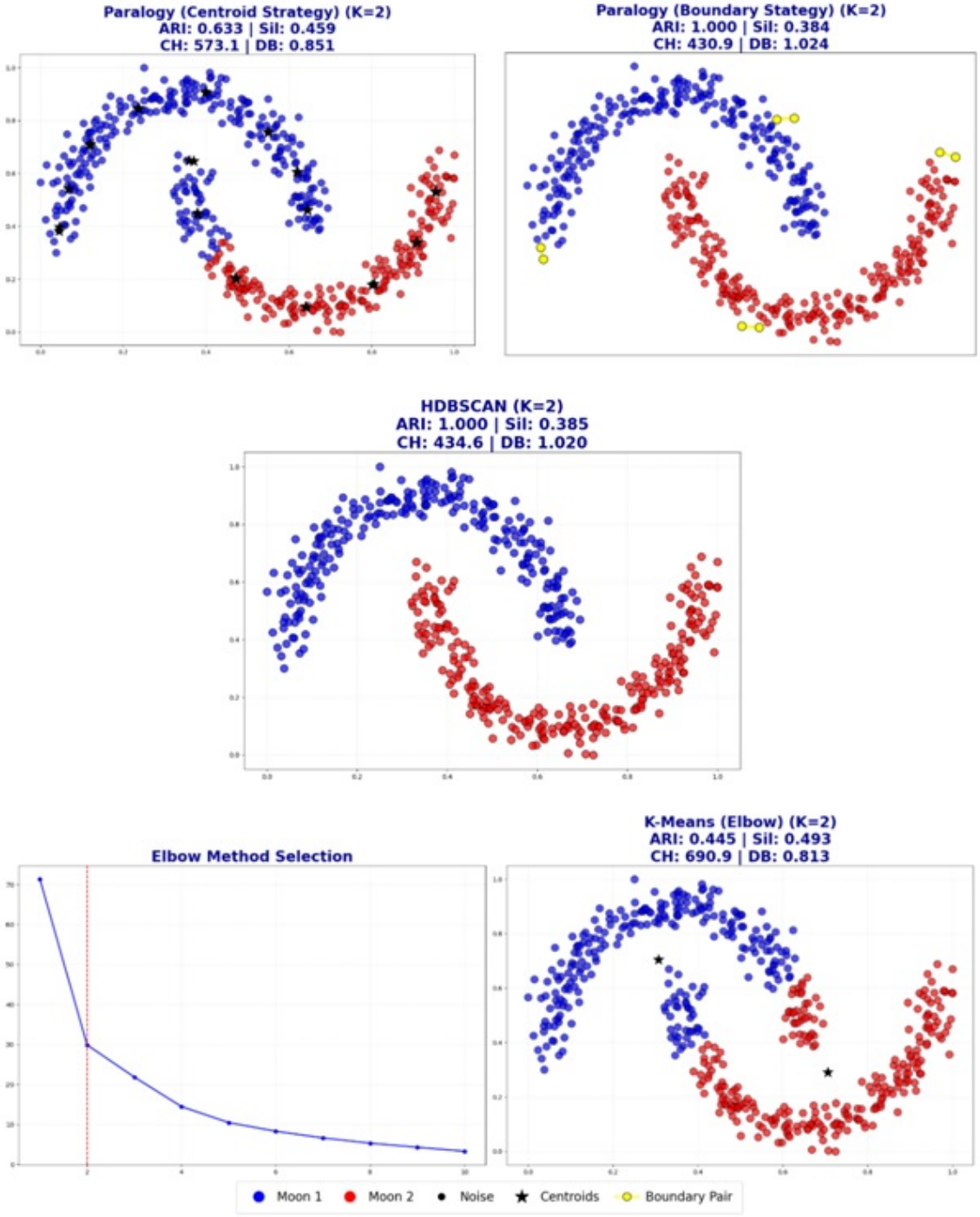


FIGURE 7 – Paralogy (stratégie centroïde)/Paralogy (stratégie frontière) /HDBSCAN/k-means : Résultats pour le jeu de données Moons

Références

[1] James C. Bezdek, Robert Ehrlich, and William Full. FCM : The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3) :191–203, 1984.

[2] Myriam Bouhnas and Henri Prade. A new approach to clustering based on an analogy-related logical proportion. In *4th Workshop Interactions between Analogical Reasoning and Machine Learning, IJCAI, 17 Aug., Montréal*. AAAI Press, 2025. Version française abrégée : Une méthode de clustering basée sur la paralogie inverse graduelle, Proc. Rencontres Francophones Logique Floue et Applications 2025, Clermont-Ferrand, Nov. 6-7, Cépaduès, 135-142.

[3] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining : Proc. 17th Pacific-Asia Conf.(PAKDD'2013), Part II*, volume 7819 of *LNCS*, pages 160–172. Springer, 2013.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96), Portland, Aug. 2-4*, page 226–231, 1996.

[5] Peter Hájek. Metamathematics of fuzzy logic. Kluwer, 1998.

[6] Anil K Jain. Data clustering : 50 years beyond k-means. *Pattern Recognition Letters*, 31(8) :651–666, 2010.

[7] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2) :129–136, 1982.

[8] Henri Prade and Gilles Richard. Reasoning with logical proportions. In Fangzhen Lin, Ulrike Sattler, and Miroslaw Truszczynski, editors, *Proc. 12th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2010), Toronto, May 9-13*. AAAI Press, 2010.

[9] Henri Prade and Gilles Richard. Homogeneous logical proportions : Their uniqueness and their role in similarity-based prediction. In Gerhard Brewka, Thomas Eiter, and Sheila A. McIlraith, editors, *Proc. 13th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2012), Rome, June 10-14*. AAAI Press, 2012.

[10] Henri Prade and Gilles Richard. From analogical proportion to logical proportions. *Logica Universalis*, 7(4) :441–505, 2013.

[11] Henri Prade and Gilles Richard. Homogenous and heterogeneous logical proportions. *IfCoLog J. of Logics and their Applications*, 1(1) :1–52, 2014.

[12] Henri Prade and Gilles Richard. Pair comparators : a family of connectives related to logical proportions. In B. Vantaggi *et al.*, editor, *Proc. 21st Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'26), Rome, June 15-19*, Communications in Computer and Information Science. Springer, 2026.

[13] David E. Rumelhart and Adele A. Abrahamson. A model for analogical reasoning. *Cogn. Psychol.*, 5 :1–28, 1973.

Vieux mais pas obsolètes : Bag-of-Words vs. Embeddings en modélisation thématique

Jean-Charles Lamirel[1,2], Francis Lareau[3], Christophe Malaterre[4], Thibault Prouteau[2]

[1]Université de Strasbourg, [2]MosAIk-LORIA, [3] Université de Sherbrooke / UQAM, [4] UQAM, CIRST

lamirel@loria.fr

Résumé

Cet article compare quatre méthodes de modélisation thématique — LDA, CFMf, Top2Vec et BERTopic — sur un corpus de 16 917 articles de philosophie des sciences, selon des critères de cohérence, de diversité et de rappel. Les résultats montrent que Top2Vec excelle en cohérence et diversité mais échoue en rappel et interprétabilité. BERTopic surpasse légèrement LDA en cohérence mais pas en rappel. CFMf, grâce à une adaptation métrique angulaire du clustering GNG, offre le meilleur équilibre entre toutes les dimensions. Ces résultats démontrent la compétitivité durable des approches BOW et soulignent la nature modulaire des pipelines de modélisation thématique.

Mots-clés

Modélisation thématique, Bag-of-Words, embeddings, LDA, CFMf, BERTopic, Top2Vec, cohérence, diversité, rappel.

Abstract

This paper compares four topic modeling approaches — LDA, CFMf, Top2Vec, and BERTopic — on a corpus of 16,917 philosophy of science articles across coherence, diversity, and recall metrics. Results show Top2Vec leads on coherence and diversity but lags on recall and interpretability; BERTopic marginally outperforms LDA on coherence; CFMf, using an angular GNG adaptation, achieves the best overall balance. Statistical BOW-based models remain competitive against modern embedding methods.

Keywords

Topic modeling, Bag-of-Words, embeddings, LDA, CFMf, BERTopic, Top2Vec, coherence, diversity, recall.

1 Introduction

La modélisation thématique est un outil important de l'analyse computationnelle de corpus textuels. Les approches classiques fondées sur la représentation Bag-of-Words (BOW) — notamment LDA [2] — des documents textuels ont longtemps dominé le domaine. L'émergence de modèles basés sur les plongements lexicaux/embeddings (Top2Vec [1], BERTopic [5]) a relancé la question de leur pertinence.

Cet article prolonge des travaux antérieurs sur CFMf [7], une méthode BOW combinant la maximisation des traits (Feature Maximization) et le clustering neuronal à base de gaz croissants (GNG). L'étude poursuit trois objectifs : (i) corriger les défauts résiduels de CFMf, (ii) étendre la comparaison à BERTopic, (iii) étudier la modularité des pipelines de modélisation thématique.

2 Méthodes comparées

LDA [2] est un modèle génératif statistique représentant chaque document comme un mélange de thèmes, chacun défini par une distribution sur les mots. Il utilise une représentation initiale BOW des documents textuels et une inférence bayésienne (échantillonnage de Gibbs ou inférence variationnelle).

CFMf [7] utilise aussi une représentation BOW des documents et combine la maximisation des traits (F-mesures de traits) (6] pour le ranking des mots avec le clustering GNG [4] pour le regroupement. GNG est un algorithme winner-take-most adaptatif qui apprend la topologie des données sans fixer le nombre de clusters a priori. Une version angulaire de GNG est introduite dans ce travail : à chaque étape d'apprentissage, les vecteurs prototypes des clusters gagnants sont renormalisés après la mise à jour des poids. Cette renormalisation projette les prototypes sur la sphère unité, rendant la compétition entre neurones (prototypes) insensible au nombre de mots présents dans les documents et évitant ainsi les clusters attracteurs [7].

Top2Vec [1] utilise Doc2Vec pour créer des plongements sémantiques denses de mots et documents. HDBSCAN fait émerger des clusters denses sans fixer K a priori ; les centroïdes de clusters servent de représentants thématiques et le ranking des mots repose sur leur proximité cosinus au centroïde.

BERTopic [5] exploite des plongements contextuels profonds (ici réalisés avec le modèle Transformer stella_en_1.5B_v5), HDBSCAN pour le clustering et la réassignation des outliers, et une pondération c-TFIDF pour le ranking des mots. BERTopic illustre ainsi la modularité des pipelines en combinant des plongements Transformer (pour le clustering) et une représentation BOW (pour le ranking des top-mots).

3 Protocole expérimental

Le corpus comprend 16 917 articles en texte intégral issus de huit revues de référence en philosophie des sciences (1930–2017) [8]. Après tokenisation, étiquetage morphosyntaxique et lemmatisation (TreeTagger), seuls les noms, verbes, adverbes et adjectifs apparaissant dans au moins 50 phrases sont conservés. Les modèles sont évalués pour $K \in [5, 100]$ selon quatre métriques complémentaires :

– Cohérence C_V [9] : co-occurrence des mots d'un thème ;

– Diversité thématique : ratio de mots-clés uniques sur le total des mots-clés ;

– Rappel interne micro (mIR) : capacité moyenne des mots-clés à rappeler les documents de leur thème ;

– Rappel interne joint micro (mJIR) : couverture conjointe de l'ensemble du corpus par tous les mots-clés.

Les thèmes LDA sont alignés sur les clusters de CFMf, Top2Vec et BERTopic en affectant leurs documents de manière stricte (selon leur thème dominant), ceci pour faciliter la comparaison qualitative.

4 Résultats

La **cohérence** croît avec K pour tous les modèles, atteignant un plateau vers K=20–30 pour LDA, CFMf et Top2Vec, et plus tardivement pour BERTopic. Top2Vec atteint les valeurs les plus élevées ($C_V \approx 0{,}80$ dès K=20), suivi de CFMf ($C_V > 0{,}70$). LDA affiche les scores les plus bas (plateau ≈

0,55), légèrement dépassé par BERTopic aux valeurs élevées de K.

La **diversité thématique** décroît à mesure que K augmente. Top2Vec maintient la diversité la plus haute (> 0,95 pour tout K). CFMf se classe deuxième (0,80–0,90). LDA et BERTopic convergent vers 0,65 à partir de K=30.

Le **rappel** révèle un tableau radicalement différent. Top2Vec enregistre les scores mIR les plus faibles (< 0,35), révélant une déconnexion entre ses mots-clés et ses documents. LDA domine (mIR > 0,80), suivi de BERTopic (0,70–0,80) et de CFMf (≈ 0,60 pour K > 25). Il faut noter que la valeur mIR présente une signification particulière pour détecter les modèles dont la couverture documentaire est faible, comme c'est le cas pour Top2Vec, lorsqu'elle est nettement basse. En revanche, l'interprétation de valeurs mIR élevées comme indicateurs de performances supérieures doit être abordée avec prudence. De telles valeurs élevées peuvent résulter de la présence de termes fréquents mais sémantiquement peu informatifs dans les top-mots des topics, ce qui nuit à l'interprétation de ces derniers.

Pour le mJIR, LDA, BERTopic et CFMf atteignent tous ≈ 1, indiquant une couverture quasi-totale du corpus. Top2Vec plafonne à 0,90, laissant ≈ 10 % des documents non rappelés quelle que soit la valeur de K.

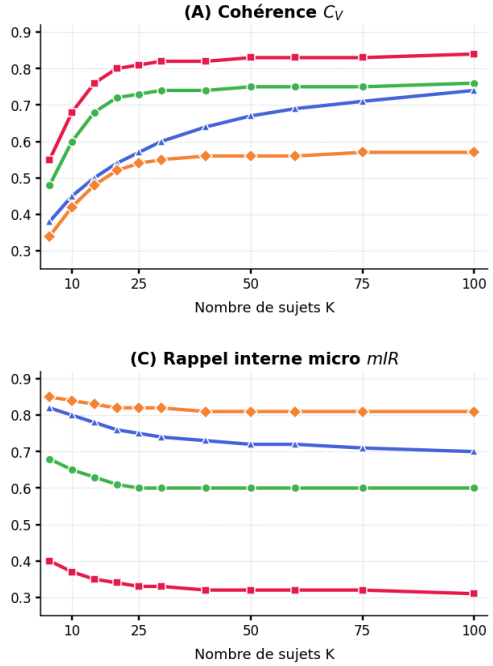


Fig. 1. Comparaison des modèles en fonction de K : (A) cohérence CV, (C) rappel interne micro mIR. Modèles : Top2Vec (carrés rouges), CFMf (cercles verts), BERTopic (triangles bleus), LDA (losanges oranges).

4.3 Interprétabilité qualitative

L'alignement thématique (K=25) montre une bonne convergence descriptive globale. Néanmoins, les top-mots thématiques de Top2Vec incluent fréquemment des noms d'auteurs et des termes très spécifiques, réduisant leur interprétabilité. CFMf génère des descripteurs précis et bien délimités — distinguant par exemple relativité et mécanique quantique — avec quelques noms d'auteurs résiduels. BERTopic produit un mélange de descripteurs clairs et de descripteurs génériques. LDA tend à produire un nombre important de descripteurs génériques

5 Discussion

Les résultats mettent en évidence des compromis fondamentaux entre approches. Les modèles à plongements (Top2Vec, BERTopic) tirent parti de représentations sémantiques riches mais présentent des faiblesses importantes en rappel pour Top2Vec, en cohérence pour BERTopic, et en interprétabilité pour les deux. Les approches BOW, en particulier CFMf, offrent un meilleur compromis vis-à-vis de la couverture, de la cohérence et de l'interprétabilité. Un aspect clé concerne la modularité des pipelines : les composantes de prétraitement, vectorisation, clustering et ranking de mots sont en grande partie indépendantes et combinables. L'adaptation angulaire de CFMf illustre comment une modification ciblée d'un seul composant peut éliminer un défaut structurel sans modifier le reste du pipeline. Des travaux futurs exploreront des stratégies de réassignation des données marginales plus fines, susceptibles d'améliorer davantage les petites classes, ainsi que la prise en charge d'embeddings par le clustering GNG dans la chaine CFMf. Parmi les limitations : le corpus est restreint à un seul domaine disciplinaire ; de nombreuses autres approches restent à explorer. Ce travail met également en évidence les limites des métriques automatiques pour mesurer la qualité des thèmes, y compris les plus couramment utilisées, comme la cohérence |9]. Cela suggère l'intérêt de les compléter par une évaluation humaine et une évaluation fondée sur les LLM, dans le cadre d'un protocole d'évaluation mixte et parallèle.

6 Conclusion

Aucune approche ne domine uniformément sur toutes les dimensions évaluées. Top2Vec excelle en cohérence et diversité mais échoue en rappel et interprétabilité. LDA et BERTopic privilégient la couverture documentaire, mais peinent en interprétabilité et en cohérence. CFMf, grâce à son adaptation angulaire, réalise cependant le meilleur compromis global et génère des mots-clés hautement interprétables. Ces résultats réhabilitent les modèles BOW face aux méthodes à plongements et soulignent le potentiel d'approches hybrides exploitant la modularité des pipelines — une piste prometteuse pour des travaux futurs.

Références

[1] D. Angelov. Top2Vec: Distributed Representations of Topics. arXiv:2008.09470, 2020.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent Dirichlet Allocation. JMLR, 3:993–1022, 2003.

[3] J. Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers. Proc. NAACL-HLT, pp. 4171–4186, 2019.

[4] B. Fritzke. A growing neural gas network learns topologies. NIPS, 7, 1994.

[5] M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794, 2022.

[6] J.-C. Lamirel, N. Dugué, P. Cuxac. New efficient clustering quality indexes. IJCNN, pp. 3649–3657, 2016.

[7] J.-C. Lamirel, F. Lareau, C. Malaterre. CFMf topic-model: Comparison with LDA and Top2Vec. Scientometrics, 129:6387–6405, 2024.

[8] C. Malaterre, F. Lareau. The early days of contemporary philosophy of science. Synthese, 200(3):242, 2022.

[9] M. Röder, A. Both, A. Hinneburg. Exploring the Space of Topic Coherence Measures. WSDM, pp. 399–408, 2015.

Processus d'identification automatique d'articles de presse dédiés à la veille syndromique

Rosalie Nkounghawe[1], Paulin Melatagia[2], Isabelle Pieretti[3,4], Carlène Trévenec[5,6], Mathieu Roche[7,8]

[1] Université Sorbonne Paris Nord, France
[2] Université de Yaoundé I, Cameroun
[3] CIRAD, UMR PHIM, F-34398 Montpellier, France.
[4] PHIM, CIRAD, INRAE, Institut Agro, IRD, Université de Montpellier, Montpellier, France.
[5] INRAE, UMR ASTRE, F-34398 Montpellier, France.
[6] ASTRE, CIRAD, INRAE, Université de Montpellier, Montpellier, France.
[7] CIRAD, UMR TETIS, F-34398 Montpellier, France.
[8] TETIS, Université de Montpellier, AgroParisTech, CIRAD, INRAE, Montpellier, France

corine.nkounghawe@facsciences-uy1.cm, paulin.melatagia@facsciences-uy1.cm,
isabelle.pieretti@cirad.fr, carlene.trevennec@cirad.fr, mathieu.roche@cirad.fr

Résumé

Le projet SURSY (SURveillance SYndromique) vise à améliorer la détection précoce d'événements sanitaires en santé végétale à partir de sources textuelles issues du Web. L'un des principaux défis de ce domaine réside dans la rareté et l'hétérogénéité des données disponibles, rendant difficile l'apprentissage de modèles robustes. Dans ce travail, nous proposons une approche fondée sur plusieurs méthodes d'intelligence artificielle pour structurer, classifier et enrichir des données textuelles liées à la veille syndromique.

Notre contribution s'articule en trois phases. La première consiste à classifier les articles selon trois thématiques (santé animale, santé végétale, santé publique) à l'aide d'embeddings SBERT et d'un perceptron multicouches, après un équilibrage du corpus réalisé via une combinaison UMAP + K-means. La deuxième phase porte sur la détection automatique des articles de veille syndromique, en comparant une approche lexicale (TF-IDF + XGBoost) à une approche sémantique reposant sur le fine-tuning de SBERT. Enfin, une troisième phase exploratoire évalue l'impact de l'augmentation de données par rétro-traduction sur la classe minoritaire.

Mots-clés

veille syndromique, fouille de texte, traitement automatique du langage, classification de textes, évaluation des modèles

Abstract

The SURSY (SYndromic SURveillance) project aims to improve the early detection of plant health events using textual information extracted from Web sources. One of the main challenges in this domain lies in the scarcity and heterogeneity of available data, which makes it difficult to train robust models. In this work, we propose an approach based on several artificial intelligence methods to structure, classify, and enrich textual data related to syndromic surveillance.

Our contribution is organized into three phases. The first phase focuses on classifying articles into three thematic categories (animal health, plant health, public health) using SBERT embeddings and a multilayer perceptron, after balancing the corpus through a combination of UMAP and K-means clustering. The second phase addresses the automatic detection of syndromic surveillance articles by comparing a lexical approach (TF-IDF + XGBoost) with a semantic approach based on fine-tuned SBERT. Finally, a third exploratory phase evaluates the impact of data augmentation through back-translation on the minority class.

Keywords

syndromic surveillance, text mining, natural language processing, text classification, model evaluation

1 Introduction

La veille sanitaire repose de plus en plus sur l'exploitation automatique de sources d'information issues du Web afin de détecter précocement l'émergence ou la réémergence de maladies. Dans les domaines de la santé animale et végétale, plusieurs plateformes de surveillance ont été développées pour analyser des flux continus d'articles de presse et de contenus en ligne [20]. Parmi elles, ESA [1] (Epidémiosurveillance en Santé Animale) et ESV [2] (Epidémiosurveillance en Santé Végétale) constituent des dispositifs opérationnels permettant de suivre l'apparition d'événements sanitaires à partir de sources ouvertes [2]. Dans ce cadre, la plateforme PADI-web [3], dédiée à la surveillance événe-

1. https ://plateforme-esa.fr/fr
2. https ://plateforme-esv.fr/
3. https ://www.padi-web-one-health.org

mentielle en santé animale et végétale, collecte automatiquement des articles de presse multilingues et applique un filtrage thématique pour identifier des signaux sanitaires pertinents [21].

Cependant, ces dispositifs atteignent leurs limites lorsqu'il s'agit d'identifier des phénomènes émergents pour lesquels la cause est encore inconnue ou mal caractérisée. La veille syndromique vise à répondre à ce besoin en s'appuyant sur l'analyse de signaux faibles issus de descriptions de symptômes, indépendamment de l'identification explicite d'une maladie. Si cette approche est largement étudiée en santé publique et animale, notamment via des systèmes tels que HealthMap ou ProMED [3], son application à la santé végétale reste encore peu explorée, alors même que les enjeux liés aux maladies émergentes, aux nouvelles plantes hôtes et au changement climatique sont croissants.

Dans ce contexte, l'un des principaux défis de la veille syndromique en santé végétale réside dans la rareté et la diversité limitée des données textuelles disponibles, rendant difficile l'apprentissage de modèles de classification robustes. Le projet SURSY s'inscrit dans cette problématique en proposant d'étendre les systèmes de veille existants à la détection de signaux syndromiques en santé végétale, tout en consolidant les liens avec la veille en santé animale.

Dans cet article, nous proposons une approche de classification automatique de textes appliquée à des articles issus de sources Web et de la plateforme PADI-web. Le travail est structuré en trois phases principales : une première phase dédiée à la classification thématique, une seconde phase centrée sur l'identification des articles relevant de la veille syndromique, et une troisième phase exploratoire portant sur l'augmentation de données textuelles. Nous présentons les méthodes mises en œuvre, une synthèse de certaines expérimentations réalisées et les résultats obtenus, en mettant l'accent sur leur pertinence pour des applications opérationnelles de veille syndromique en santé végétale.

2 État de l'art

2.1 Veille syndromique et systèmes existants

La veille syndromique repose sur l'analyse de signaux faibles décrivant des symptômes avant la confirmation d'un diagnostic. En santé publique, plusieurs systèmes exploitent des sources ouvertes pour détecter précocement des événements émergents. Parmi eux, HealthMap [3], ProMED [4] et GPHIN [7] constituent des références majeures pour la surveillance mondiale des maladies infectieuses.

En santé animale et végétale, des plateformes telles que ESA, ESV permettent la surveillance d'événements sanitaires à partir de sources médiatiques [2]. Toutefois, ces systèmes restent principalement centrés sur des maladies connues et ne couvrent que partiellement les signaux faibles caractéristiques de la veille syndromique. À ce jour, peu de travaux portent sur la veille syndromique végétale, malgré des similarités fortes avec les descriptions de symptômes en santé animale.

2.2 Classification automatique de textes

Les approches classiques de classification textuelle reposent sur des représentations lexicales telles que TF-IDF [1], souvent combinées à des modèles supervisés comme les SVM [11] ou XGBoost [5]. Ces méthodes, bien que robustes et rapides, restent limitées par leur dépendance au vocabulaire et leur difficulté à capturer la sémantique profonde des textes.

Les modèles de langue pré-entraînés ont profondément transformé le traitement automatique du langage naturel (TALN). BERT [6], SBERT [18] et RoBERTa [14] ont démontré une capacité supérieure à représenter le sens des phrases, notamment dans des contextes où les formulations sont variées, implicites ou synonymiques. Leur adaptation par fine-tuning permet en général d'obtenir des résultats de bonne qualité même sur des corpus spécialisés.

2.3 Données rares et déséquilibrées

La classification en contexte de données déséquilibrées constitue un défi majeur, en particulier lorsque la classe d'intérêt est fortement minoritaire. Dans ce cas, les modèles supervisés ont tendance à privilégier la classe majoritaire, ce qui dégrade les performances sur la classe cible. L'utilisation de métriques adaptées est essentielle : l'AUC-PR est notamment recommandée pour évaluer les performances sur des classes minoritaires [19], car elle met l'accent sur la précision et le rappel de la classe positive.

Plusieurs stratégies permettent de limiter les effets du déséquilibre, notamment la pondération des classes, la sélection de données ou l'utilisation de représentations sémantiques plus robustes. Une synthèse des méthodes d'apprentissage sur données déséquilibrées est proposée dans [9], qui souligne l'importance de combiner des techniques d'équilibrage avec des modèles capables de capturer des structures complexes dans les données.

2.4 Augmentation de données textuelles

L'augmentation de données vise à enrichir artificiellement les classes minoritaires afin d'améliorer la robustesse des modèles. Les techniques courantes incluent la synonymie, la suppression ou l'insertion de mots, ainsi que la rétrotraduction [8], qui consiste à traduire un texte vers une langue pivot avant de le retraduire vers la langue d'origine. Ces méthodes permettent de générer des variantes textuelles tout en préservant le sens global.

Des approches plus récentes, telles que EDA (Easy Data Augmentation) [22], proposent des transformations simples mais efficaces pour diversifier les données textuelles. Par ailleurs, les modèles génératifs modernes ouvrent des perspectives pour produire des textes synthétiques plus variés et plus cohérents. Parmi eux, l'approche Retrieval-Augmented Generation (RAG) [13] combine un module de recherche d'information et un modèle génératif afin de produire des textes s'appuyant sur des documents réels. Cette méthode permet de générer des contenus plus diversifiés et plus fidèles au domaine étudié.

2.5 Contributions

Les contributions principales de ce travail sont les suivantes :

— la construction d'un corpus textuel structuré pour la veille syndromique en santé végétale, mis à disposition publiquement ;
— une approche de classification thématique (santé animale, santé végétale, santé publique) fondée sur SBERT et un équilibrage *UMAP + K-means*. UMAP est donc utilisé pour projeter les embeddings SBERT dans un espace de dimension réduite tout en préservant la structure locale des données. Cette projection facilite l'application de K-means, qui permet ensuite de sélectionner de manière contrôlée des échantillons représentatifs dans chaque cluster. Cette combinaison assure un équilibrage par thématique qui préserve la diversité sémantique du corpus tout en limitant les biais liés aux thématiques sur-représentées ;
— un modèle de détection d'articles de veille syndromique reposant sur le fine-tuning de SBERT, surpassant les approches lexicales. Plusieurs variantes de SBERT ont été expérimentées, mais une seule retenue que nous présentons dans la section 4 ;
— une exploration de l'augmentation de données par rétro-traduction et son impact sur la classe minoritaire.

3 Données

Les données utilisées dans ce travail proviennent principalement de **PADI-web**, une plateforme de veille automatisée collectant quotidiennement des articles de presse multilingues liés à la santé animale et végétale. Les corpus ont été constitués et annotés différemment selon les besoins des trois phases méthodologiques. Cette section décrit précisément leur origine, leur structure et leurs caractéristiques quantitatives.

3.1 Corpus pour la classification thématique (Phase 1)

Le premier corpus regroupe des articles issus de PADI-web, organisés initialement par *maladie* (ex. Avian Influenza, African Swine Fever, Xylella Fastidiosa) et par *type de maladies* (santé animale (SA), santé publique (SP) et santé végétale (SV)). Les fichiers fournis contenaient les colonnes *title*, *text*, *host_type*, *disease*, *year* et *relevance*.

Plusieurs opérations ont été réalisées pour pré-traiter le corpus : suppression de doublons (environ 4200 articles), correction des titres manquants, suppression des textes corrompus (détection via *langdetect* et ratio de lettres), suppression des noms de maladies, fusion des fichiers en un corpus unique.

La taille du corpus initial est répartie de la manière suivante :

— SA : 93 398 textes
— SP : 8 819 textes
— SV : 8 168 textes

La longueur des textes est résumée ci-dessous :

— Moyenne : 230 mots
— Médiane : 180 mots
— Min/Max : 40–1200 mots

Après équilibrage UMAP + K-means détaillé dans les sections suivantes, chaque classe (SA, SP, SV) contient 8168 textes.

3.2 Corpus pour la détection de la veille syndromique (Phase 2)

Ce corpus vise à distinguer les articles VS (veille syndromique) des articles NVS (non veille syndromique). Les annotations ont été réalisées manuellement par des experts du CIRAD et de l'INRAE selon un protocole interne basé sur la présence de descriptions de symptômes, l'absence de diagnostic explicite et la détection de signaux faibles.

Les 2250 articles composant le corpus sont répartis de la manière suivante :

— VS : 225 textes ($\approx 10\%$)
— NVS : 2025 textes ($\approx 90\%$)

La longueur des textes est résumée ci-dessous :

— Moyenne : 260 mots
— Médiane : 210 mots
— Min/Max : 50–1500 mots

3.3 Annotations thématique et pertinence

Les annotations de données pour ce projet ont été réalisées par des expertes du domaine qui sont co-autrices de cet article et spécialisées en veille épidémiologique au sein des plateformes ESV[4] et ESA[5]. Un protocole multi-annotateurs avec mesure d'accord inter-annotateurs est envisagée dans nos futurs travaux.

3.4 Corpus augmenté (Phase 3)

Pour enrichir la classe minoritaire VS, une augmentation par rétro-traduction (anglais $\rightarrow$ allemand $\rightarrow$ anglais) a été appliquée et détaillée dans la section suivante.

La taille du corpus après augmentation est :

— NVS : 2241 textes
— VS : 498 textes après suppression des doublons une fois l'augmentation réalisée.

La longueur moyenne des textes augmentés est de 270 mots, avec une variabilité légèrement accrue due aux reformulations.

4 Méthodologie

La méthodologie adoptée dans ce travail est structurée en trois phases principales, correspondant aux objectifs du projet SURSY. Chaque phase produit un jeu de données spécifique, mis à disposition publiquement [15]. Le code

4. La Plateforme ESV (Plateforme d'Épidémiosurveillance en Santé Végétale) est en charge d'améliorer la surveillance de la santé des végétaux en France en associant 7 acteurs publics et privés nationaux reconnus en leur haut niveau d'expertise.

5. La Plateforme nationale d'Épidémiosurveillance en Santé Animale apporte un appui méthodologique et opérationnel aux services compétents de l'État et aux autres gestionnaires de dispositifs de surveillance pour la conception, le déploiement, l'animation, la valorisation et l'évaluation des dispositifs de surveillance sanitaire et biologique du territoire.

associé à l'ensemble des expérimentations est disponible en accès libre [17]. L'ensemble du pipeline méthodologique s'appuie à la fois sur des approches lexicales classiques et sur des modèles de langage pré-entraînés, conformément aux recommandations de travaux récents en classification de textes [6, 18].

Protocole d'évaluation : Le protocole d'évaluation appliqué dans les trois phases est adapté à la nature de chaque modèle. Pour les classifieurs ne nécessitant pas de sélection de modèle en cours d'entraînement (Random Forest, Naive Bayes, régression logistique), nous utilisons une séparation *train*/*test* stratifiée, avec une graine aléatoire fixée pour la reproductibilité. Pour les modèles entraînés de façon itérative avec arrêt anticipé ou sélection du meilleur *checkpoint* (MLP, XGBoost avec arrêt anticipé, et fine-tuning de modèles pré-entraînés tels que SBERT ou RoBERTa), nous utilisons une séparation *train*/*validation*/*test* stratifiée, où l'ensemble de validation sert à la sélection du meilleur modèle pendant l'entraînement et l'ensemble de test est réservé à l'évaluation finale. Dans les deux cas, la stratification préserve la distribution des classes entre les ensembles, ce qui est particulièrement important pour la Phase 2 où la classe VS est fortement minoritaire. Pour garantir la comparabilité entre les approches évaluées au sein d'une même phase, les indices des ensembles sont sauvegardés et réutilisés à l'identique pour chaque modèle comparé. Les métriques rapportées (accuracy, F1-score, AUC multi-classes pour la Phase 1, AUC-PR pour la Phase 2) sont choisies en fonction de la nature de la tâche et du déséquilibre éventuel des classes.

4.1 Phase 1 : Classification des articles par thématique

Cette première phase vise à structurer les données textuelles selon trois thématiques : santé publique (SP), santé animale (SA) et santé végétale (SV). Cette étape permet de réduire les biais liés à des sources ou maladies sur-représentées et constitue un prérequis essentiel pour la veille syndromique.

Prétraitement et vectorisation : Les articles issus de PADI-web sont nettoyés (suppression des balises HTML, normalisation, suppression des doublons). Deux types de représentations textuelles sont explorés : TF-IDF, couramment utilisé dans les approches lexicales [11], et SBERT (`all-MiniLM-L6-v2`), modèle de plongement sémantique performant pour la classification [18]. Cette version de SBERT a été retenue comme vectorizer dans ce projet pour son bon compromis entre qualité des représentations sémantiques et coût de calcul : six couches Transformer, 22M de paramètres, et des performances compétitives sur les tâches de similarité sémantique [18].

Équilibrage des données : La distribution initiale des thématiques (SA, SP, SV) étant fortement déséquilibrée, un équilibrage est nécessaire pour éviter que le modèle n'apprenne des biais structurels. Nous avons comparé deux stratégies :

- **Sous-échantillonnage par maladie** : limitation du nombre d'articles par couple thématique × maladie, afin de garantir une représentation équivalente de chaque maladie au sein de chaque thématique.
- **Combinaison UMAP + K-means** : UMAP (Uniform Manifold Approximation and Projection) réduit la dimensionnalité des embeddings SBERT tout en préservant la structure locale des données [16] ; K-means est ensuite appliqué dans l'espace réduit pour identifier des clusters cohérents, et un nombre équilibré d'articles est sélectionné dans chaque cluster afin de préserver la diversité sémantique tout en homogénéisant la distribution entre SA, SV et SP.

La stratégie d'équilibrage **UMAP + K-means** est particulièrement adaptée aux corpus textuels hétérogènes, comme le montrent des travaux similaires en classification thématique [1].

Ce type de combinaison « réduction de dimension + clustering » pour l'équilibrage de corpus textuels déséquilibrés est proche de travaux récents. Comme dans les travaux de Last *et al.* [12], qui proposent un ré-échantillonnage guidé par K-means (K-Means SMOTE) afin de mieux exploiter la structure du corpus, nous utilisons K-means pour regrouper les articles en zones sémantiques cohérentes avant d'opérer une sélection équilibrée. De manière similaire, Jamil *et al.* [10] ont récemment évalué l'association entre UMAP et des méthodes de rééquilibrage pour la classification de textes médicaux, confirmant l'intérêt d'une projection manifold-learning en amont de la gestion du déséquilibre. Notre approche s'inscrit dans cette lignée, tout en se distinguant par le fait que nous ne générons pas d'échantillons synthétiques : nous sélectionnons uniquement des articles réels, ce qui garantit la plausibilité linguistique des textes retenus. Cette stratégie évite les écueils bien connus de SMOTE dans le cadre textuel, où la génération d'échantillons synthétiques par interpolation dans l'espace des embeddings ne correspond pas nécessairement à des textes linguistiquement valides.

Modélisation de notre solution : Les embeddings SBERT équilibrés servent d'entrée à un perceptron multicouches (MLP) composé de deux couches denses avec activation ReLU. L'entraînement utilise un arrêt anticipé basé sur la perte de validation. Les performances sont évaluées via l'accuracy, le F1-score et l'AUC multi-classes.

Comparaison des approches : Pour justifier les choix méthodologiques décrits ci-dessus (combinaison UMAP + K-means pour l'équilibrage et SBERT + MLP pour la classification), nous avons croisé les deux stratégies d'équilibrage avec trois combinaisons vectorisation/classifieur. Les résultats obtenus sur l'ensemble de test sont rassemblés dans le Tableau 1. Les performances sont systématiquement meilleures avec UMAP + K-means qu'avec le sous-échantillonnage par maladie, et SBERT + MLP s'est révélé le plus stable sur des phrases courtes ou des formulations s'écartant du vocabulaire d'entraînement, ce qui a motivé notre choix final.

Équilibrage	Approche	Acc.	F1	AUC
Par maladie	TF-IDF + RF	0.92	0.92	0.98
Par maladie	SBERT + RF	0.93	0.93	0.98
Par maladie	SBERT + MLP	0.93	0.93	0.99
UMAP + K-means	TF-IDF + RF	0.98	0.98	1.00
UMAP + K-means	SBERT + RF	0.97	0.97	1.00
UMAP + K-means	SBERT + MLP *(retenu)*	**0.98**	**0.98**	**1.00**

TABLE 1 – Comparaison des approches testées pour la classification thématique (Phase 1), sur l'ensemble de test. F1 et AUC sont calculés en moyenne macro (*one-vs-rest* pour l'AUC sur les trois thématiques SA, SP, SV). La combinaison UMAP + K-means avec SBERT + MLP a été retenue pour son comportement plus stable sur des formulations n'appartenant pas strictement au vocabulaire d'entraînement.

4.2 Phase 2 : Détection des articles de veille syndromique

La deuxième phase constitue le cœur du projet et vise à distinguer les articles de veille syndromique (VS) des articles non veille syndromiques (NVS). Cette tâche est formulée comme une classification binaire fortement déséquilibrée (environ 10% VS).

Prétraitement et vectorisation : Les textes sont normalisés puis vectorisés selon deux approches complémentaires :

- **TF-IDF + XGBoost**, approche rapide et assez robuste pour des tests préliminaires, mais limitée par sa dépendance au vocabulaire [5].
- **SBERT fine-tuné**, permettant de capturer des formulations implicites et des synonymes, particulièrement utiles dans la veille syndromique où les signaux sont rarement explicites.

Plusieurs combinaisons vectorisation/classifieur ont été évaluées à titre comparatif avant d'arrêter le choix du modèle SBERT `all-MiniLM-L6-v2` ; les résultats obtenus sur l'ensemble de test pour la classe minoritaire VS sont résumés dans le Tableau 2. Deux constats se dégagent : d'une part, les approches hybrides SBERT + classifieur superficiel exploitent mal la richesse des embeddings sur un corpus aussi déséquilibré ; d'autre part, TF-IDF + XGBoost obtient de bons scores mais reste très sensible aux synonymes, comme le confirment les tests qualitatifs sur phrases courtes (voir Section 5). Le fine-tuning complet de SBERT, où tous les poids du modèle sont mis à jour via la classe `Trainer` de Hugging Face, est l'approche retenue pour le déploiement. RoBERTa a également été fine-tuné à titre exploratoire, avec des performances proches (F1-macro $\approx$ 0.93), mais pour un coût d'entraînement plus élevé qui ne justifiait pas son adoption.

4.3 Phase 3 : Augmentation de données textuelles

La troisième phase explore l'augmentation de données pour enrichir la classe VS. L'objectif est d'améliorer la robustesse du modèle face à la variabilité lexicale.

Rétro-traduction : Nous utilisons la rétro-traduction (anglais $\rightarrow$ allemand $\rightarrow$ anglais), technique couramment employée pour générer des paraphrases [8]. Afin de choisir la langue pivot, nous avons comparé deux configurations : l'anglais $\rightarrow$ français $\rightarrow$ anglais et l'anglais $\rightarrow$ allemand $\rightarrow$ anglais. Les deux configurations améliorent les performances par rapport au modèle de la Phase 2 sans augmentation, mais l'allemand offre un F1-score légèrement supérieur sur la classe minoritaire VS (voir Section 5), et a donc été retenu comme langue pivot. Cette méthode permet d'introduire des variations lexicales tout en conservant le sens.

Réentraînement et évaluation : Les textes augmentés sont intégrés au corpus d'entraînement du modèle SBERT. L'impact est mesuré via l'évolution de l'AUC-PR et la stabilité des prédictions sur des phrases tests.

Approche	Préc. (VS)	Rapp. (VS)	F1 (VS)	AUC-PR (VS)
TF-IDF + Naive Bayes	0.25	0.06	0.10	0.80
TF-IDF + Rég. logistique	0.96	0.55	0.70	0.80
SBERT + Rég. logistique	0.52	0.83	0.64	0.72
SBERT + XGBoost	0.86	0.52	0.65	0.81
TF-IDF + XGBoost	1.00	0.87	0.93	0.98
SBERT *fine-tuné* (retenu)	**0.95**	**0.87**	**0.91**	**0.95**

TABLE 2 – Comparaison des approches testées pour la détection des articles de veille syndromique (Phase 2), sur le sous-ensemble de test (225 articles, dont 23 VS et 202 NVS). Les valeurs reportées correspondent à la précision, au rappel, au F1-score et à l'AUC-PR de la classe minoritaire VS. Le fine-tuning complet de SBERT est l'approche retenue pour sa robustesse aux reformulations, confirmée par les tests qualitatifs (Section 5).

5 Résultats

Cette section présente les principaux résultats obtenus lors des différentes phases du projet, en cohérence avec la méthodologie décrite précédemment. Les résultats sont présentés de manière synthétique, en mettant l'accent sur ceux qui ont été retenus durant les expérimentations.

5.1 Résultats de la classification thématique

Pour la classification thématique des articles (santé publique, santé animale et santé végétale), nous avons retenu l'approche ayant fourni les meilleures performances, fondée sur l'apprentissage par transfert en utilisant les données équilibrées avec UMAP et K-means. Les textes sont d'abord représentés à l'aide du modèle SBERT *all-MiniLM-L6-v2*, puis les embeddings obtenus sont utilisés comme entrée d'un modèle séquentiel léger de type perceptron multicouches.

Le modèle converge après 67 époques (Figure 1a), avec des performances élevées et stables sur les ensembles d'entraînement et de validation. Les scores d'accuracy et de F1-score dépassent 97% sur l'ensemble de validation. L'évaluation sur l'ensemble de test confirme ces résultats, avec une AUC multi-classes de 0.99 (Figure 1b) et une très faible confusion entre les thématiques (Figure 1c).

Enfin, un test de prédiction sur une phrase courte (*"A new disease affecting animals was discovered this year."*)

montre une probabilité de 0.98 pour la classe correcte (Figure 1d), confirmant la robustesse du modèle.
L'ensemble de ces résultats est synthétisé en Figure 1.

5.2 Résultats de la détection des articles de veille syndromique

L'objectif de cette deuxième phase est de classifier automatiquement les articles selon leur type : veille syndromique (VS) ou non veille syndromique (NVS). Cette tâche est rendue particulièrement complexe par le fort déséquilibre des données, reflétant la réalité opérationnelle, avec une proportion d'environ 10% d'articles VS contre 90% d'articles NVS.
Parmi les multiples approches abordées durant cette phase, nous avons retenu deux méthodes, bien qu'une seule soit véritablement fiable. Nous les avons retenues pour des raisons particulières :

- **TF-IDF + XGBOOST** (Figure 2) : vectorisation avec TF-IDF et classification avec XGBoost. Cette approche est rapide en termes de calcul et donne d'assez bonnes performances, mais elle reste limitée par l'effet statique du vectorizer, très sensible aux synonymes qui ne figurent pas dans son vocabulaire.
- **Fine-tuning de SBERT** (Figure 3) : après un temps d'entraînement plus conséquent, cette approche produit un excellent rapport de classification ainsi qu'un score AUC-PR de 0.95. De plus, le meilleur modèle sauvegardé parvient à bien classer les phrases contenant des synonymes ou des formulations variées liées à la veille syndromique, ce qui n'était pas le cas de TF-IDF + XGBoost.

Nous avons conservé les deux approches car, pour un test rapide, la première est suffisante, tandis que la seconde est plus adaptée à un déploiement opérationnel nécessitant davantage de robustesse.
Des tests complémentaires réalisés sur des phrases courtes confirment que le fine-tuning de SBERT offre une meilleure capacité de généralisation. Contrairement aux modèles basés sur TF-IDF, le modèle fine-tuné sur SBERT parvient à identifier des articles relevant de la veille syndromique même lorsque les expressions employées reposent sur des synonymes ou des formulations implicites. Cette capacité à capturer le sens global des textes constitue un atout majeur pour la veille syndromique, où les signaux recherchés sont rarement explicitement formulés.
Le détail des performances des deux approches est présenté dans les Figures 2 (TF-IDF + XGBoost) et 3 (SBERT fine-tuné).

5.3 Résultats de l'augmentation de données textuelles

La troisième phase du projet a pour objectif d'explorer l'impact de l'augmentation de données textuelles sur la détection des articles de veille syndromique. Cette étape vise à enrichir la classe minoritaire (VS), fortement sous-représentée, afin d'améliorer la robustesse et la stabilité des modèles de classification.

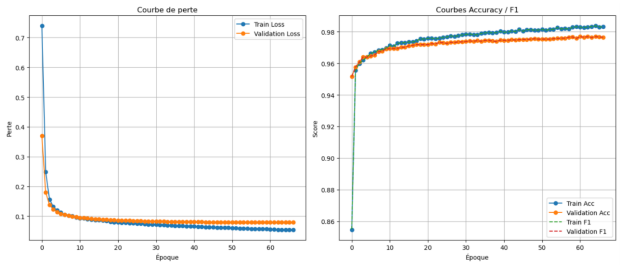
Configuration	AUC-PR	F1 (VS)
Sans augmentation (Phase 2)	0.95	0.91
Augmentation, pivot français	0.97	0.91
Augmentation, pivot allemand *(retenu)*	**0.98**	0.91

TABLE 3 – Comparaison de l'effet de la langue pivot pour la rétro-traduction (Phase 3), évalué via le fine-tuning de SBERT sur l'ensemble de test. Les deux configurations améliorent les performances par rapport à la Phase 2 sans augmentation, l'allemand offrant l'AUC-PR le plus élevé.

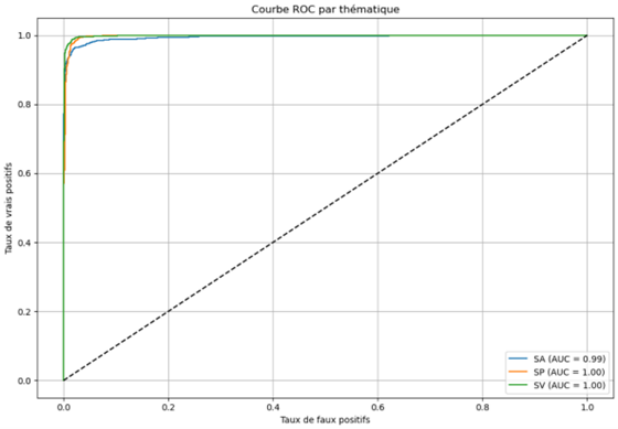
Après avoir obtenu le jeu de données par la méthode de rétro-traduction, le modèle SBERT a été réentraîné sur ces derniers via le fine-tuning. Nous avons comparé deux langues pivots : le français et l'allemand (Tableau 3). Les deux configurations améliorent les performances par rapport au modèle de la Phase 2 sans augmentation et augmentent globalement les probabilités de prédiction pour la classe VS, ce qui confirme l'apport de l'augmentation. L'allemand offre en particulier l'AUC-PR le plus élevé (0.98 contre 0.97 avec le français, voir Figure 4a), et a donc été retenu pour la suite. Le détail des métriques par classe sur l'ensemble de test est présenté en Figure 4b. Les tests réalisés sur des phrases courtes confirment une meilleure assurance du modèle dans ses décisions, suggérant un effet positif de l'augmentation sur la capacité de généralisation.
Ces premiers résultats indiquent que l'augmentation de données textuelles constitue une piste prometteuse pour renforcer la détection des articles de veille syndromique. Cette phase reste néanmoins exploratoire et ouvre des perspectives vers l'utilisation de techniques plus avancées, telles que l'exploitation de plusieurs langues pivots ou le recours à des modèles génératifs récents comme le RAG.
L'ensemble des résultats de cette phase est synthétisé en Figure 4.

6 Discussion

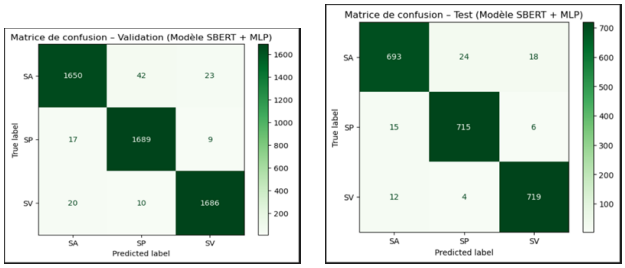
Les résultats obtenus au cours des différentes phases du projet mettent en évidence plusieurs enseignements importants pour le développement de systèmes de veille syndromique en santé végétale.
Tout d'abord, la phase de classification thématique confirme l'intérêt de structurer les données en amont afin de réduire les biais liés à des maladies, des sources ou des contextes dominants. L'utilisation de représentations sémantiques basées sur SBERT, combinée à une stratégie d'équilibrage des données (UMAP et K-means), permet d'obtenir des performances élevées et stables. Cette étape apparaît essentielle pour garantir une meilleure généralisation des modèles dans un contexte multi-domaines, et constitue un prérequis solide pour les phases ultérieures de veille syndromique.
La comparaison menée lors de la phase de détection des articles de veille syndromique met clairement en évidence les limites des approches purement lexicales, telles que TF-IDF combiné à XGBoost. Bien que rapides et efficaces pour des tests préliminaires, ces méthodes restent fortement dépendantes du vocabulaire observé lors de l'apprentissage



(a) Courbes d'apprentissage (loss / accuracy).



(b) Courbes ROC multi-classes (one-vs-rest).

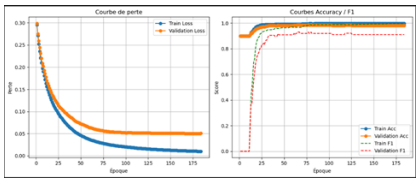


(c) Matrices de confusion : validation (gauche) et test (droite).

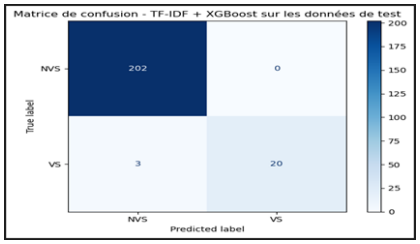
Texte	*A new disease affecting animals was discovered this year*
Thématique prédite	SA
Score de confiance	0.9870
Probabilités	
SA	0.9870
SP	0.0102
SV	0.0028

(d) Exemple de prédiction.

FIGURE 1 – Synthèse des résultats de la phase 1 (classification thématique).



(a) Courbe d'apprentissage (TF-IDF + XGBoost).

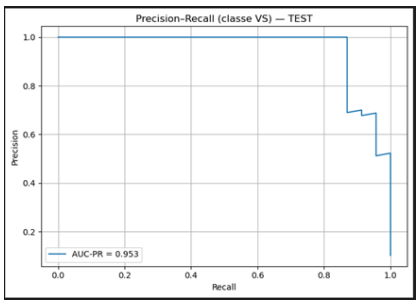


(b) Matrice de confusion (test).

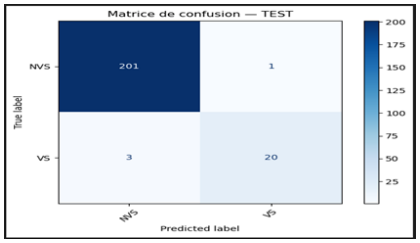
Classe	Prec.	Rapp.	F1	Supp.
NVS	0.99	1.00	0.99	202
VS	1.00	0.87	0.93	23
Acc.			0.99	
Macro	0.99	0.91	0.96	225
Weighted	0.99	0.99	0.99	225

(c) Synthèse des métriques (TF-IDF + XGBoost).

FIGURE 2 – Résultats TF-IDF + XGBoost : (a) courbe d'apprentissage, (b) matrice de confusion, (c) synthèse des métriques.



(a) PR (classe VS).



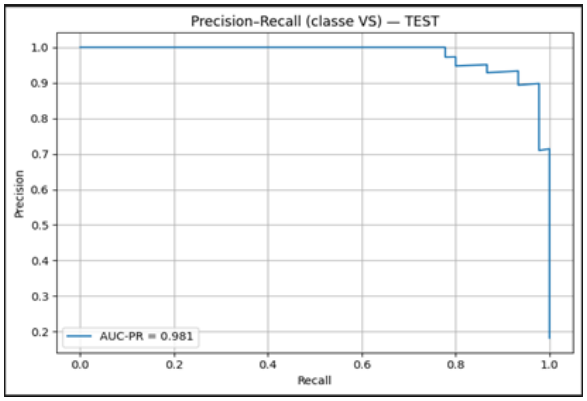
(b) Confusion (test).

Classe	Prec.	Rapp.	F1	Supp.
NVS	0.99	1.00	0.99	202
VS	0.95	0.87	0.91	23
Acc.			0.98	
Macro	0.97	0.93	0.95	225
Weighted	0.98	0.98	0.98	225

(c) Synthèse métriques.

FIGURE 3 – Résultats de SBERT fine-tuné.

et ont des difficultés à capturer des formulations implicites ou synonymiques. À l'inverse, les approches sémantiques reposant sur le fine-tuning de modèles de langage pré-entraînés, comme SBERT, montrent une meilleure capacité à saisir le sens global des textes et à détecter des signaux faibles, caractéristique essentielle dans un contexte



(a) Courbe Precision–Recall (classe VS).

Classe	Prec.	Rapp.	F1	Supp.
NVS	0.98	0.99	0.98	202
VS	0.93	0.89	0.91	45
Acc.		0.97		
Macro	0.95	0.94	0.94	247
Weighted	0.97	0.97	0.97	247

(b) Synthèse des métriques (test).

FIGURE 4 – Résultats de la phase 3 : (a) courbe Precision–Recall sur la classe VS et (b) synthèse des métriques sur l'ensemble de test, après augmentation par rétro-traduction (allemand comme langue pivot). L'augmentation conduit à un test set contenant 45 articles VS (contre 23 en Phase 2 sans augmentation).

de veille syndromique.

Néanmoins, la détection des articles de veille syndromique demeure une tâche complexe en raison du fort déséquilibre entre les classes. Les résultats obtenus confirment que ce déséquilibre constitue un frein majeur à l'apprentissage de modèles robustes, en particulier pour la classe minoritaire. Bien que les représentations sémantiques améliorent la cohérence et la stabilité des prédictions, elles restent sensibles à la quantité limitée de données disponibles en santé végétale.

Les premiers résultats issus de la phase exploratoire d'augmentation de données textuelles suggèrent que l'enrichissement artificiel de la classe minoritaire constitue une piste prometteuse. L'augmentation par rétro-traduction introduit une variabilité lexicale supplémentaire qui se traduit par une amélioration modérée des performances et une meilleure assurance du modèle dans ses prédictions. Toutefois, cette approche reste limitée par la simplicité des transformations appliquées et devra être approfondie afin de générer des textes plus diversifiés et plus représentatifs des situations réelles de veille.

Enfin, des expérimentations complémentaires ont été réalisées pour tester les capacités de transfert des modèles en apprenant sur des données en SA (resp. SV) et testant en SV (resp. SA). Le Tableau 4 montre des résultats satisfaisants quant aux capacités de transfert des modèles. Ceci peut s'expliquer par un vocabulaire utilisé qui peut être relativement commun pour décrire des évenements syndromiques. Ceci met en évidence l'intérêt de renforcer les liens entre la veille en santé animale, santé publique et la veille en santé végétale. Les similarités observées dans les descriptions de symptômes ouvrent des perspectives intéressantes pour le transfert de connaissances entre domaines, notamment dans le contexte de maladies émergentes ou de phénomènes de type « maladie X » (maladies inconnues), dont les manifestations peuvent évoluer rapidement sous l'effet du changement climatique.

Train : SV / Test : SA	F1
NVS	1
VS	0.96
Train : SA / Test : SV	**F1**
NVS	0.98
VS	0.81

TABLE 4 – Résultats de transfert fondés sur la méthode TF-IDF+XGBOOST (notons que les noms de maladie ont été supprimés comme pré-traitement pour éviter de potentiels bias).

7 Conclusion et perspectives

Ce travail mené dans le cadre du projet SURSY [6] a exploré l'apport des modèles de langue pour la détection automatique d'articles de veille syndromique en santé végétale, dans un contexte marqué par des données rares, hétérogènes et fortement déséquilibrées. Les résultats obtenus montrent que les approches sémantiques, et en particulier les modèles SBERT fine-tunés, sont mieux adaptées que les méthodes lexicales classiques pour capturer des signaux faibles et implicites caractéristiques de la veille syndromique.
L'étude met également en évidence l'intérêt de l'augmentation de données textuelles par rétro-traduction pour renforcer la robustesse des modèles, notamment vis-à-vis de la classe minoritaire. Bien que les gains observés restent modérés, cette approche contribue à améliorer la stabilité des prédictions et la confiance du modèle.
Les perspectives de ce travail portent sur l'exploration de techniques d'augmentation plus avancées, incluant l'usage de plusieurs langues pivots ainsi que l'intégration de modèles génératifs récents, tels que les approches basées sur le RAG. À plus long terme, l'extension de ces méthodes à des scénarios de veille multi-domaines, en lien avec la santé animale et publique (contexte One Health), constitue une voie prometteuse pour renforcer la détection précoce de maladies émergentes dans un contexte de changement

6. https ://clapas.umontpellier.fr/projets/projet-sursy/

climatique et de mobilité humaine.

Remerciements

Les auteurs tiennent à remercier les bailleurs de fonds CLAPAS[7] (Collaborative Local Research Actions on Plant health and AgroSystems) – Université de Montpellier, la DGAL (Direction générale de l'alimentation) ainsi que le projet ANR BEYOND (20-PCPA-0002) dont le soutien a contribué au bon déroulement de ce travail. Enfin, les auteurs remercient Sarah Valentin (Cirad, TETIS) qui a mis à disposition un corpus dédié la veille syndromique.

Références

[1] Charu C Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In *Mining Text Data*, pages 163–222. Springer, 2012.

[2] Elena Arsevska, Mathieu Roche, Pascal Hendrikx, and Renaud Lancelot. From syndromic surveillance to event-based surveillance : monitoring animal health events using online news. *Preventive Veterinary Medicine*, 160 :135–144, 2018.

[3] John S Brownstein, Clark C Freifeld, Ben Y Reis, and Kenneth D Mandl. Healthmap : global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2) :150–157, 2008.

[4] Martin Carrion and Lawrence C Madoff. Promedmail : 22 years of digital disease detection. *Online Journal of Public Health Informatics*, 9(1), 2017.

[5] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[7] Marc Dion, Peter AbdelMalik, and Abla Mawudeku. Gphin : Global public health intelligence network. *Online Journal of Public Health Informatics*, 7(1), 2015.

[8] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of EMNLP*, pages 489–500, 2018.

[9] Haibo He and Edwardo Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9) :1263–1284, 2009.

[10] Arslan Jamil, Muhammad Kashif Hanif, Muhammad Umer Sarwar, and Muhammad Irfan Khan. An empirical evaluation of dimensionality reduction and class balancing for medical text classification. *Scientific Reports*, 16(815), 2026.

[11] Thorsten Joachims. Text categorization with support vector machines : Learning with many relevant features. In *Proceedings of ECML*, pages 137–142, 1998.

[12] Felix Last, Georgios Douzas, and Fernando Bacao. Oversampling for imbalanced learning based on K-Means and SMOTE. *arXiv preprint arXiv :1711.00837*, 2017.

[13] Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

[14] Yinhan Liu et al. Roberta : A robustly optimized bert pretraining approach. In *arXiv preprint arXiv :1907.11692*, 2019.

[15] Nkounghawe Rosalie ; Meltatagia Paulin ; Pieretti Isabelle ; Trevenec Carlène ; Roche Mathieu. Sursy data : Textual data for syndromic surveillance. https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/MYMPSO, 2025. Dataverse.

[16] Leland McInnes, John Healy, and James Melville. Umap : Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv :1802.03426*, 2018.

[17] Rosalie NKOUNGHAWE. Sursy : Projet_sursy. https://github.com/rosalie0309/Projet_SURSY, 2025.

[18] Nils Reimers and Iryna Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992, 2019.

[19] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. In *PLOS ONE*, volume 10, page e0118432. Public Library of Science, 2015.

[20] Sarah Valentin, Elena Arsevska, Alize Mercier, Sylvain Falala, Julien Rabatel, Renaud Lancelot, and Mathieu Roche. Padi-web : An event-based surveillance system for detecting, classifying and processing online news. In Zygmunt Vetulani, Patrick Paroubek, and Marek Kubis, editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 87–101, Cham, 2020. Springer International Publishing.

[21] Sarah Valentin, Elena Arsevska, Julien Rabatel, Sylvain Falala, Alizé Mercier, Renaud Lancelot, and Mathieu Roche. Padi-web 3.0 : A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health*, 13 :100357, 2021.

[22] Jason Wei and Kai Zou. Eda : Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of EMNLP*, 2019.

7. https ://clapas.umontpellier.fr/

Prédiction du délai de vente des véhicules et optimisation du stock par apprentissage automatique

Hamid Ahaggach[1], Lylia Abrouk,[2], Eric Lebon [1]

[2] Laboratoire LIB , Université Bourgogne Europe, Dijon, France
prenom.nom@u-bourgogne.fr
[1] Syartec, Aix-en-Provence, France
hahaggach@Syartec.com, elebon@Syartec.com

Résumé

La gestion des stocks automobiles nécessite d'anticiper le délai de vente des véhicules afin de limiter les coûts d'immobilisation et la dépréciation. Cet article présente la démonstration d'un système d'aide à la décision fondé sur l'apprentissage automatique pour prédire le délai de vente d'un véhicule à partir de ses caractéristiques techniques et commerciales, puis recommander des décisions d'approvisionnement. Le problème est formulé comme une tâche de classification multi-classes, où le délai de vente est discrétisé en intervalles temporels. Le système proposé intègre un pipeline complet de prétraitement, un module de prédiction supervisée et un module décisionnel permettant d'estimer un stock cible compatible avec un objectif de rotation. La démonstration met en évidence l'intégration opérationnelle du modèle prédictif dans une interface interactive permettant d'explorer des scénarios de type what-if *pour la gestion proactive des stocks.*

Mots-clés

Classification supervisée, Sélection de caractéristiques, Optimisation décisionnelle ; Système d'aide à la décision ; Random Forest ; Analyse prédictive ; Gestion de stock automobile.

Abstract

Automotive inventory management requires anticipating vehicle time-to-sale in order to reduce holding costs and depreciation. This paper presents the demonstration of a machine-learning-based decision-support system designed to predict vehicle sales delay from technical and commercial attributes, and to support procurement decisions. The problem is formulated as a multi-class classification task, where time-to-sale is discretized into time intervals. The proposed system integrates a complete pipeline including data preprocessing, supervised prediction, and a decision module that estimates stock levels compatible with a target sales horizon. The demonstration highlights the operational integration of predictive models into an interactive application that enables users to explore what-if scenarios for proactive inventory management.

Keywords

Supervised classification ; Feature selection ; Decision optimization ; Decision support system ; Random Forest ; Predictive analytics ; Automotive inventory management.

1 Introduction

La rotation de stock constitue un enjeu central pour les concessionnaires et les plateformes de revente [4] : un véhicule immobilisé trop longtemps génère des coûts (financement, stockage, assurance), augmente le risque de décote et dégrade la trésorerie. À l'inverse, des stocks trop faibles entraînent des ruptures et des pertes d'opportunité commerciale. Dans ce contexte, anticiper le délai de vente d'un véhicule, dès son entrée en stock ou lors d'une décision d'achat, permet d'arbitrer entre disponibilité, marge et risque. Des travaux récents montrent que l'analyse prédictive et l'apprentissage automatique peuvent améliorer les décisions d'inventaire et réduire les coûts opérationnels dans l'industrie automobile [6, 8, 5].

Les délais de vente dépendent d'un ensemble de facteurs hétérogènes (marque, modèle, version, motorisation, couleur, saisonnalité, historique de stock), et les données opérationnelles sont souvent incomplètes, irrégulières et non stationnaires (évolution des préférences, politiques commerciales, contraintes d'approvisionnement). L'intégration de modèles d'intelligence artificielle et d'analytique avancée permet d'exploiter ces sources d'information hétérogènes afin d'améliorer la précision des prévisions et la gestion des stocks [3, 1].

Les approches de séries temporelles (ARIMA, Prophet, modèles séquentiels) supposent généralement des observations régulières et des historiques suffisamment longs par référence (produit, modèle, version). Or, dans les stocks automobiles, l'historique par référence fine (marque–modèle–version–options) est limité et irrégulier. Beaucoup de références apparaissent sporadiquement, rendant l'apprentissage de motifs temporels instable, et les ruptures d'approvisionnement introduisent des périodes sans observation. Des études récentes soulignent que dans ces contextes à demande intermittente, les méthodes d'apprentissage supervisé sur données tabulaires peuvent offrir une meilleure robustesse que les modèles temporels classiques [7, 2]. En

pratique, le problème se prête mieux à une formulation tabulaire supervisée : on prédit une étiquette de délai à partir de variables explicatives décrivant le véhicule et son contexte (date, stock, etc.). Cette approche permet : (i) d'exploiter des données multi-références dans un modèle global, (ii) d'intégrer naturellement des variables catégorielles, (iii) de produire des recommandations actionnables.

2 Problème et approche

On considère un concessionnaire qui achète et revend des véhicules. Chaque véhicule est décrit par un vecteur de caractéristiques x regroupant des informations techniques, commerciales et contextuelles (marque, modèle, version, motorisation, couleur, date d'estimation, niveau de stock, etc.). L'objectif est de prédire le délai de vente y, discrétisé en plusieurs classes :

$$y \in \{\texttt{<3 m}, \texttt{3–6 m}, \texttt{6–9 m}, \texttt{9–12 m}, \texttt{>12 m}\}.$$

Le premier objectif est donc de résoudre une tâche de classification supervisée :

$$\hat{y} = f_\theta(x),$$

où f_θ désigne un classifieur appris à partir de données historiques.

Le second objectif est décisionnel. Pour une référence donnée, le concessionnaire souhaite estimer un stock maximal recommandé $n^\star$ compatible avec un objectif de rotation, par exemple rester dans la classe `<3 mois`. Le principe consiste à faire varier la variable de stock dans le vecteur d'entrée afin d'identifier le seuil à partir duquel la classe prédite change.

La figure 1 présente l'architecture générale du système de démonstration. Celui-ci est structuré en trois composantes : un module de prétraitement des données, un module de prédiction du délai de vente, et un module d'optimisation décisionnelle. Les données manipulées sont de nature tabulaire et comprennent des variables catégorielles (marque, modèle, version, motorisation, couleur), des variables temporelles dérivées de la date d'estimation, ainsi qu'une variable d'action correspondant au nombre d'exemplaires en stock. Le prétraitement assure la validation des entrées, la construction de variables temporelles, l'encodage des variables catégorielles et la gestion des incohérences ou valeurs manquantes.

Le module de prédiction applique un classifieur supervisé au vecteur encodé. Plusieurs familles de modèles peuvent être envisagées, notamment les arbres de décision, les forêts aléatoires ou les méthodes d'ensemble. Dans notre prototype, le choix d'une Random Forest offre un compromis satisfaisant entre robustesse, simplicité d'usage et capacité à modéliser des interactions non linéaires entre variables.

Le module décisionnel transforme ensuite la prédiction en recommandation d'approvisionnement. Pour une référence donnée, le système simule différentes valeurs de stock et observe l'évolution de la classe prédite. Une stratégie en deux temps est appliquée : une recherche grossière pour localiser une zone de bascule, puis une recherche fine pour estimer précisément le seuil de réapprovisionnement. Le système peut ainsi fournir soit un nombre maximal de véhicules à acheter, soit une recommandation d'attente avant réapprovisionnement.

3 Démonstration du système

Afin d'illustrer l'intégration opérationnelle de l'approche proposée, nous avons développé un prototype interactif permettant d'explorer les prédictions et les recommandations de stock. L'objectif de la démonstration est de montrer comment un utilisateur métier peut renseigner les caractéristiques d'un véhicule, obtenir une estimation du délai de vente et analyser différents scénarios de réapprovisionnement.

L'interface du prototype a été conçue pour un usage décisionnel. Elle permet de saisir les principales caractéristiques d'un véhicule, d'afficher le délai de vente prédit et de consulter les recommandations associées. L'application propose également des vues complémentaires, notamment une vue de prédiction et une vue dédiée aux recommandations ou véhicules similaires. La restitution des résultats s'appuie sur une hiérarchie visuelle simple combinant zones de paramétrage, cartes de synthèse et tableaux interactifs. Cette organisation facilite l'exploration rapide des résultats par des utilisateurs non spécialistes de l'apprentissage automatique.

Lors de la démonstration, l'utilisateur sélectionne une configuration de véhicule à partir de ses caractéristiques principales et du stock courant. Le système retourne alors une classe de délai de vente prédite, accompagnée d'une recommandation de réapprovisionnement. L'utilisateur peut ensuite modifier la valeur du stock pour explorer des scénarios de type *what-if* et observer l'évolution de la recommandation. Ce scénario met en évidence l'intérêt du système pour l'aide à la décision dans un contexte de gestion proactive des stocks.

Le prototype est implémenté sous la forme d'une application web interactive (Figure 2). Afin de garantir une expérience fluide lors de la démonstration, le système met en cache les ressources du modèle ainsi que certains calculs intermédiaires. Cette stratégie limite les recalculs inutiles et permet de conserver les résultats lors des interactions successives avec l'interface.

4 Évaluation

L'évaluation du système repose sur un jeu de données opérationnel fourni par deux concessionnaires automobiles couvrant la période d'octobre 2013 à novembre 2023. L'ensemble de données comprend plus de 73 200 véhicules décrits par 34 attributs techniques et commerciaux, incluant notamment la marque, le modèle, la motorisation, l'année, ainsi que diverses informations liées au stock et à la commercialisation. Les données proviennent de deux entreprises distinctes et incluent à la fois des véhicules neufs (VN) et des véhicules d'occasion (VO) [4].

Les données ont été divisées aléatoirement en deux ensembles : 80% pour l'entraînement et 20% pour l'évalua-

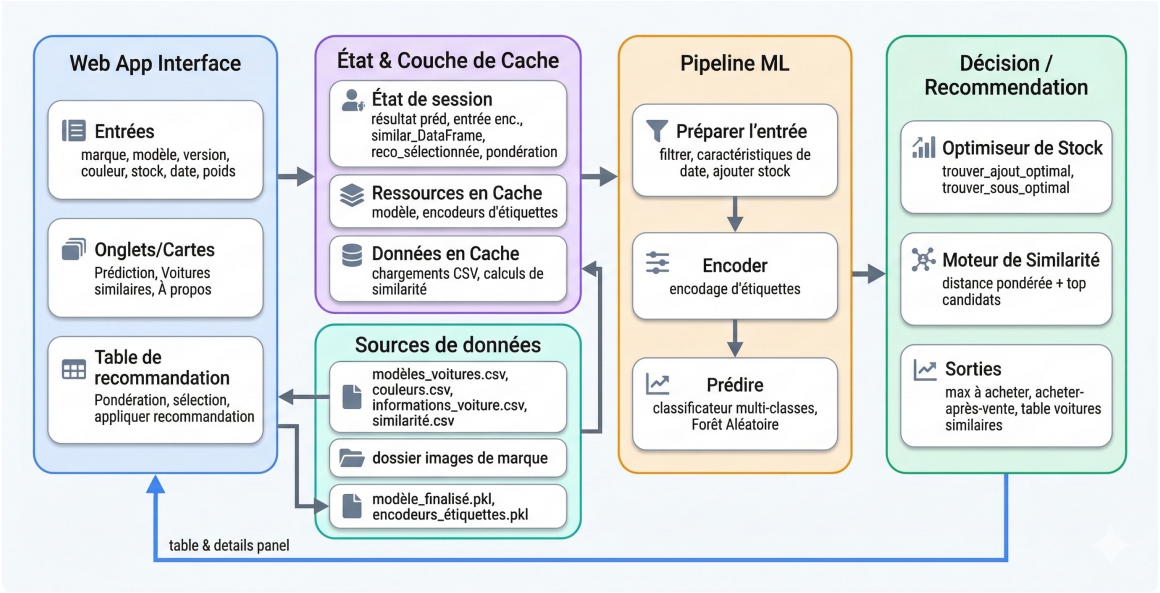


FIGURE 1 – Architecture générale du système de démonstration.

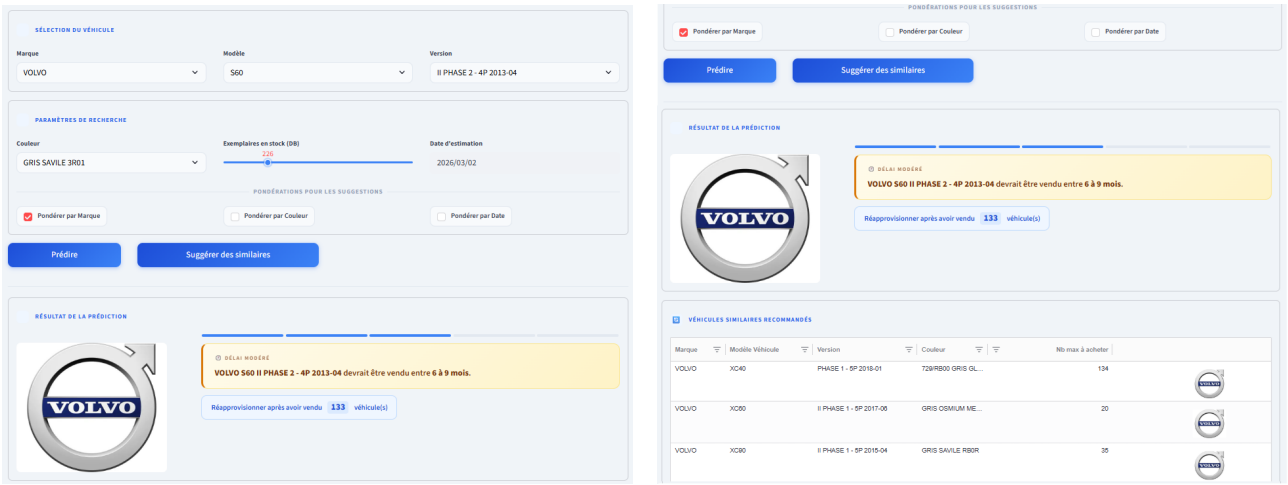


FIGURE 2 – Vues du prototype de démonstration : interface de prédiction et vue de recommandation.

tion. L'objectif du modèle est de prédire l'intervalle de temps pendant lequel un véhicule restera en stock avant d'être vendu. Les délais de vente sont discrétisés en plusieurs classes correspondant à des intervalles temporels (par exemple : `0–3 mois`, `3–6 mois`, `6–9 mois`, `9–12 mois`, `>12 mois`).

Plusieurs algorithmes d'apprentissage supervisé ont été évalués, notamment KNN, SVM, Decision Tree et Random Forest. Les hyperparamètres ont été optimisés à l'aide d'une recherche par grille (*GridSearch*) afin d'obtenir les meilleures performances possibles pour chaque modèle.

Les résultats expérimentaux montrent que le modèle Random Forest obtient les meilleures performances globales. Pour le premier concessionnaire, le modèle atteint une précision de 99.53% pour les véhicules neufs et 92.09% pour les véhicules d'occasion. Des résultats similaires sont observés pour le second concessionnaire, avec une précision de 99.24% pour les véhicules neufs et 91.80% pour les véhicules d'occasion. Ces performances s'expliquent en grande partie par la capacité des forêts aléatoires à combiner plusieurs arbres de décision et à capturer des relations non linéaires complexes dans les données.

Les matrices de confusion montrent que la majorité des véhicules sont correctement classés dans leur intervalle de délai de vente. Les erreurs observées concernent principalement les classes adjacentes, notamment entre les intervalles `0–3 mois` et `3–6 mois`. Ce phénomène est cohérent avec la nature du problème : deux véhicules très similaires peuvent être vendus à quelques jours d'intervalle et être ainsi associés à des classes différentes.

Ces résultats confirment la pertinence de l'approche proposée pour la prédiction du délai de vente. Ils montrent également que les modèles d'ensemble, tels que Random Forest, constituent une solution robuste pour exploiter des données

tabulaires hétérogènes issues de contextes opérationnels.

5 Discussion

Les résultats expérimentaux mettent en évidence plusieurs éléments importants concernant la prédiction du délai de vente dans le secteur automobile. Tout d'abord, les performances obtenues montrent que les approches d'apprentissage supervisé sur données tabulaires sont particulièrement adaptées à ce type de problème. Contrairement aux modèles de séries temporelles classiques, qui nécessitent des historiques longs et réguliers par référence, les modèles tabulaires peuvent exploiter simultanément les caractéristiques de nombreux véhicules et généraliser plus facilement à de nouvelles configurations.

L'utilisation d'un modèle Random Forest s'avère particulièrement pertinente dans ce contexte. Les forêts aléatoires permettent de capturer des interactions complexes entre variables, tout en restant relativement robustes au bruit et aux données manquantes. Dans notre cas, elles offrent un bon compromis entre précision, stabilité et temps d'entraînement.

L'analyse des matrices de confusion révèle cependant certaines limites. Les erreurs de classification apparaissent principalement entre classes temporelles proches. Par exemple, un véhicule vendu en trois mois et un jour est classé différemment d'un véhicule vendu en deux mois et vingt-neuf jours, bien que ces situations soient très proches d'un point de vue métier. Ce phénomène est inhérent à la discrétisation du délai de vente en classes.

Une autre limite concerne la variabilité des marchés automobiles. Les préférences des consommateurs, les politiques commerciales ou les conditions économiques peuvent évoluer rapidement, ce qui peut entraîner une dérive des distributions de données au cours du temps. Dans un contexte opérationnel, il est donc nécessaire de réentraîner régulièrement les modèles afin de maintenir leur performance.

Malgré ces limites, les résultats montrent que l'intégration d'un modèle prédictif dans un système interactif peut constituer un outil d'aide à la décision pertinent pour les concessionnaires. En combinant prédiction et simulation de scénarios, il devient possible d'explorer différents niveaux de stock et d'évaluer leur impact potentiel sur la rotation des véhicules.

6 Conclusion

Dans cet article, nous avons présenté la démonstration d'un système d'aide à la décision dédié à la prédiction du délai de vente et à l'optimisation des stocks automobiles. L'approche proposée repose sur un modèle d'apprentissage supervisé exploitant les caractéristiques techniques et commerciales des véhicules afin d'estimer leur délai de vente sous forme d'intervalles temporels.

Les expérimentations réalisées sur un jeu de données réel comprenant plus de 73 000 véhicules ont montré que les modèles d'ensemble, en particulier Random Forest, permettent d'obtenir des performances élevées pour cette tâche. Les résultats confirment que l'utilisation de données tabulaires hétérogènes constitue une approche efficace pour modéliser les dynamiques de vente dans le secteur automobile.

Au-delà de la performance prédictive, l'un des apports principaux de ce travail réside dans l'intégration du modèle dans un prototype interactif permettant d'explorer des scénarios décisionnels. L'utilisateur peut ainsi analyser l'impact de différents niveaux de stock sur le délai de vente prédit et obtenir des recommandations opérationnelles pour la gestion des inventaires.

Ces résultats ouvrent plusieurs perspectives de recherche. À court terme, l'intégration de variables économiques supplémentaires telles que le prix, les promotions ou la marge commerciale pourrait améliorer la qualité des prédictions. À plus long terme, des approches hybrides combinant apprentissage automatique et connaissances métier, par exemple sous forme d'ontologies ou de règles décisionnelles, pourraient permettre d'améliorer l'interprétabilité des recommandations et leur adoption dans des contextes industriels.

Références

[1] Advancements in sales forecasting : A critical evaluation of machine learning approaches. *International Journal of Engineering Trends and Technology*, 2024.

[2] Demand forecasting in the automotive industry : A systematic literature review. *MDPI Systems*, 2024.

[3] Machine learning-based demand and pricing optimization for a resilient automotive supply chain. *IJRASET*, 2024.

[4] Hamid Ahaggach, Lylia Abrouk, Sebti Foufou, and Eric Lebon. Predicting car sale time with data analytics and machine learning. In *IFIP International Conference on Product Lifecycle Management*, pages 399–409. Springer, 2022.

[5] So Fukuhara, Abdallah Alabdallah, Nuwan Gunasekara, and Slawomir Nowaczyk. Bridging forecast accuracy and inventory kpis : A simulation-based software framework. *arXiv preprint arXiv :2601.21844*, 2026.

[6] Pranav et al. Gaikwad. Predictive logistics management of car sales based on machine learning algorithm for supply chain. 2024.

[7] Cornelius Hake, Jonas Weigele, Frederik Reichert, and Christian Friedrich. Evaluation of artificial intelligence methods for lead time prediction in automotive production. *arXiv preprint arXiv :2501.07317*, 2025.

[8] Manoj Omprakash. Optimizing demand forecasting and inventory management with ai in the automotive industry. Master's thesis, LUT University, 2024.

SESSION 3 : IA, Éthique et Enseignement présidée par Anna Pappa

Indicateurs émotionnels et posturaux pour l'intervention pédagogique précoce : modéliser la difficulté comme une déviation de la réussite individuelle de l'étudiant.

Edouard Nadaud[1,2], Lionel Prevost[1,2], Bénédicte Le Grand[2], Bastien Tarot[1,3]

[1] ESIEA, ESIEALAB-LDR (Learning Data Robotics)

[2] Paris 1 Panthéon Sorbonne, CRI (Centre de Recherche en Informatiques)

[3] ESSCA School of Management

[1]prénom.nom@esiea.fr [2]prénom.[Le-]nom@univ-paris1.fr [3]Prénom.NOM@essca.fr

Résumé

Face à la « cécité attentionnelle » inhérente à l'enseignement hybride, nous présentons SEPT, un cadre "privacy-by-design" analysant les trajectoires émotionnelles et physiques captées pendant les sessions de travail des étudiants via une webcam. En modélisant la difficulté comme une déviation de la signature personnelle de réussite plutôt que par des normes de groupe, notre approche améliore la détection de l'échec +17%. Le système permet une alerte précoce robuste dès les premiers tiers de l'activité, sans compromis significatif sur la performance. Respectueux de la vie privée, le cadre SEPT a des performances équivalentes aux méthodes biométriques intrusives pour identifier efficacement les 10-15% d'étudiants en détresse.

Mots-clés

Informatique affective, trajectoires émotionnelles, applications d'apprentissage automatique, IA éthique, analyse de l'apprentissage, mouvements du visage.

Abstract

To address the "attentional blindness" inherent in blended learning, we present SEPT, a privacy-by-design framework that analyzes learners' emotional and physical trajectories from webcam feeds. By modeling difficulty not as a deviation from a group norm, but as a departure from the student's individual "signature of success," our approach improves detection by 17%. The system enables robust early warning within the first 30% of activity. Completely privacy-respecting, SEPT achieves performance equivalent to intrusive biometric methods (AUC 0.67) in effectively identifying students at risk of dropping out.

Keywords

Affective Computing, Emotion Trajectories, Machine-Learning Applications, Ethical AI, Learning analytics, Head movements.

1 Introduction

Dans les environnements d'apprentissage à distance et les environnements d'apprentissage à distance et hybrides, les enseignants sont confrontés à une « cécité attentionnelle » car il leur est impossible de suivre les signaux non verbaux d'une cohorte entière pour identifier les étudiants en détresse. Si l'informatique affective offre des solutions automatisées pour restaurer ce lien pédagogique, le domaine se heurte à deux obstacles majeurs.

Premièrement, de nombreux modèles agrègent les états émotionnels sur une session (ex. « frustration moyenne »), écrasant ainsi la dimension temporelle. Mathématiquement, ces approches échouent à distinguer la « difficulté productive » (une confusion transitoire menant à la compréhension) de la « frustration destructrice » (une difficulté persistante menant au désengagement), car ces deux états peuvent produire des effets et des états mentaux identiques ou proches.

Deuxièmement, les avancées récentes en Apprentissage Profond, telles que les *Transformers* ou les LSTM, modélisent certes ces dynamiques temporelles, mais introduisent une barrière de « soif de données » (Data Hunger). Ces architectures exigent des jeux de données annotés massifs pour éviter le surapprentissage, ce qui les rend inapplicables dans les contextes de classe typiques (« *Small Data* »). De plus, elles reposent souvent sur un traitement vidéo brut, soulevant de graves préoccupations éthiques au regard du RGPD [1] ou l'AI ACT européen [2].

L'objectif de cette étude est de concilier une analyse temporelle granulaire avec des contraintes strictes de confidentialité. Nous proposons un cadre *privacy-by-design* exploitant la dynamique des trajectoires comportementales pour permettre une détection robuste et précoce de la difficulté académique.

L'article est organisé comme suit : la section 2 sera consacrée à l'état de l'art. La section 3 présentera nos principales contributions méthodologiques. Dans la section 4 seront détaillés nos résultats expérimentaux. La section 5 conclura nos travaux et sera suivie d'une discussion et de perspectives en section 6.

2 État de l'art

2.1 Informatique Affective et Engagement

La Théorie Contrôle-Valeur [3] établit que les émotions d'accomplissement modulent l'attention et l'effort cognitif [4]. Ces états sont généralement repérés via le modèle

dimensionnel de Russell [5] (Valence/Arousal). Toutefois, la réalité pédagogique est plus nuancée : des états a priori négatifs comme la confusion peuvent être bénéfiques s'ils viennent à être résolus [4]. Bien que les méta-analyses confirment la corrélation entre succès et émotion positive [3], la réalité" est plus nuancée. La confusion, bien que négative, peut être productive si elle est transitoire et résolue (problème de l'impasse cognitive) ; lorsqu'elle persiste, elle conduit au désengagement.

Dans ce contexte, le concept d'engagement en tant qu'objet multidimensionnel combinant des aspects comportementaux (effort, persistance), émotionnels (intérêt, valeur) et cognitifs (profondeur de traitement) devient central [6]. Les émotions agissent comme les antécédents immédiats de cet engagement : un étudiant peut ressentir de la frustration (émotion) mais maintenir son effort (engagement comportemental) jusqu'à un point de rupture. Cependant, les méthodes traditionnelles pour identifier ces états souffrent de limites significatives : s'appuyer sur l'auto-évaluation par questionnaire a posteriori ne permet pas de passer à l'échelle, tandis que les capteurs de signaux physiologiques comme l'EEG sont intrusifs. Le défi actuel est donc d'inférer ces états internes complexes uniquement à partir de signaux externes observables sans perturber les activités académiques.

2.2 Représentation de l'Engagement en Informatique

Pour relever ce défi, l'analyse de la dynamique temporelle de l'apprentissage nécessite des descripteurs continus (comme la valence et l'arousal) capables de discriminer les états affectifs. Cette approche est ancrée dans la théorie de la Cognition Incarnée (*Embodied Cognition*) [7, 8], qui postule que les processus cognitifs, tels que la charge mentale ou la résolution de problèmes, se manifestent par des ajustements moteurs subtils et involontaires. Dans ce contexte, les mouvements physiques, particulièrement ceux de la tête et du regard, sont critiques.

Bien que le regard soit unanimement reconnu comme le marqueur le plus direct pour le foyer visuel de l'attention [9], l'oculométrie traditionnelle impose des contraintes de calibration et matérielles incompatibles avec un déploiement écologique à grande échelle. Pour répondre à ces limites, l'Estimation de la Pose de la Tête (HPE) a émergé comme une alternative robuste. Des études comparatives récentes démontrent une forte corrélation entre l'orientation de la tête et la direction du regard dans des contextes d'interaction humain-écran, validant l'utilisation de l'HPE comme un estimateur fiable de l'attention en conditions naturelles [10].

Les travaux en analyse de données d'apprentissage ont identifié des rôles sémantiques distincts pour les différents degrés de liberté de la tête [11] :

- Lacet (*Yaw*) : La recherche utilisant des réseaux de neurones convolutifs a établi que la variance de l'angle de lacet discrimine efficacement entre la concentration sur le contenu pédagogique et les distractions environnementales (« vagabondage mental »), offrant une mesure quantitative de la persistance de l'attention [12].
- Roulis (*Roll*) : Des études plus récentes sur la communication non verbale associent l'inclinaison de la tête à un marqueur d'intérêt ou de curiosité, mais aussi, en cas de persistance, à de la confusion ou à une tentative de résoudre une dissonance cognitive [13]. Contrairement aux mouvements de navigation, un roulis signale souvent un changement interne de perspective face à une information ambiguë [14].

Ainsi, la littérature suggère que la combinaison des dynamiques émotionnelles (expressions faciales) et des micromouvements de la tête (HPE) offre une représentation de l'engagement plus riche que l'un ou l'autre canal pris isolément [10].

2.3 Des dynamiques temporelles à la reconnaissance précoce

Bien que l'informatique affective ait fait des progrès significatifs dans la détection d'états émotionnel [15], les modèles pédagogiques prédominants souffrent encore d'un « angle mort temporel ». De nombreuses approches actuelles reposent largement sur l'agrégation statistique, telle que la « valence moyenne » à l'échelle de la session ou la fréquence des émotions basiques d'Ekman pour prédire la performance [16]. Ce lissage statistique est mathématiquement réducteur : une valence moyenne, « neutre », est identique pour les deux profils radicalement opposés de l'étudiant totalement désengagé et de celui oscillant entre une frustration productive et des moments de compréhension [13, 17].

Les limites de l'agrégation sont bien documentées. Une étude majeure analysant les émotions d'élèves de collège n'a trouvé aucune différence significative dans le temps moyen passé dans la confusion ou l'ennui entre les élèves performants et ceux en difficulté [18]. L'absence de corrélation robuste entre les moyennes émotionnelles simples et la performance académique suggère que ce n'est pas la présence d'une émotion qui détermine le résultat, mais sa dynamique. C'est la persistance (durée) et la transition (séquence) des états qui sont prédictives : une brève frustration suivie d'un réengagement est bénigne [19] ; une frustration persistante est délétère. Par conséquent, le défi n'est plus de classifier ce que l'étudiant ressent à l'instant *t*, mais de comprendre comment sa trajectoire émotionnelle évolue dans le temps.

En suivant l'évolution pas à pas des états affectifs, on capture les dynamiques affectives : les émotions surgissent, se transforment ou s'estompent continuellement durant l'apprentissage. Cette vue dynamique s'aligne avec les théories modernes de l'affect et de la cognition, qui postulent une interaction entre émotions, engagement et processus cognitifs dans le temps [20]. Ainsi, connaître la succession des émotions (et pas seulement leur moyenne) donne un aperçu de la manière dont l'apprenant réagit aux défis cognitifs, les régule ou subit leur impact. Les théories de la régulation émotionnelle soutiennent cette vue : la capacité d'un apprenant à gérer la frustration, surmonter la confusion ou maintenir l'engagement se déploie nécessairement dans le

temps à travers une série d'ajustements.
Ces constats motivent la modélisation des signaux affectifs et comportementaux comme des séries temporelles multivariées.

2.4 Modélisation de Séries Temporelles : Vers la Reconnaissance Précoce d'Action

La trajectoire émotionnelle adopte une approche inspirée de la Reconnaissance Précoce d'Action (*Early Action Recognition*) en vision par ordinateur. Dans ce paradigme, l'objectif est de classifier une séquence (par ex., un geste humain) dès ses premières images, sans attendre son achèvement. Transposé à l'éducation, nous postulons qu'une « trajectoire d'échec » (caractérisée, par exemple, par une hésitation prolongée couplée à une désorganisation du regard) possède une signature temporelle identifiable bien avant la validation finale de la réponse par l'étudiant.
De plus, les approches actuelles en Apprentissage Profond (Transformers, LSTM) tente de résoudre cela en modélisant des séquences, mais elles introduisent une nouvelle barrière : la « Voracité en Données ». Ces architectures nécessitent des jeux de données annotés massifs qui les rendant inapplicables dans les contextes de classe typiques de « Small Data » (taille de la cohorte inférieure à 100 individus) où les contraintes RGPD limitent la collecte de données [12]. Contrairement aux actions physiques discrètes, les états cognitifs internes comme la « confusion » manquent de vérité terrain visuelle non ambiguë. Définir un étalon-or (gold standard) nécessite soit une annotation experte coûteuse, qui souffre d'une faible fiabilité inter-juges due à l'interprétation subjective des états, soit des auto-évaluations des étudiants, qui perturbent leurs activités et introduisent un biais.
Nous utilisons ici des mesures de similarité élastique, spécifiquement la DTW (*Dynamic Time Warping*) [21]. Contrairement à la distance Euclidienne, qui exige un alignement temporel rigide, la DTW permet l'alignement non linéaire de deux séries en minimisant les effets de distorsion temporelle. Cette propriété est cruciale pour modéliser le rythme cognitif : deux étudiants peuvent suivre le même chemin intellectuel (lecture → réflexion → résolution) mais à des vitesses différentes. En traitant l'évolution des signaux multimodaux comme une forme dynamique plutôt que comme une séquence d'états indépendants, la DTW permet la quantification de la similarité structurelle entre une tentative en cours et des trajectoires typiques, facilitant ainsi une classification précoce et explicable.
Cette étude poursuit deux objectifs de recherche principaux :

- Proposer une approche robuste et respectueuse de la vie privée pour identifier les signes de difficulté académique en utilisant des données non intrusives et conformes à la confidentialité.
- Valider le potentiel du système pour l'alerte précoce, permettant un soutien opportun avant que l'échec ne survienne.

Nous démontrons que cette approche atteint une utilité d'alerte précoce comparable aux méthodes invasives de l'état de l'art, mais avec une architecture entièrement respectueuse de la vie privée adaptée au déploiement dans le monde réel.

3 Proposition et méthodologie

Pour aller au-delà de l'état de l'art, nous proposons le cadre SEPT (Figure 1) afin de répondre aux questions de recherche ouvertes identifiées ci-dessus. Cela nous permet de modéliser les dynamiques temporelles de la difficulté sans compromettre la vie privée de l'étudiant. Nous postulons que l'étudiant possède une signature cinématique multimodale durant son parcours d'apprentissage, et que la difficulté se manifeste comme une signature de déviation identifiable par rapport à cette dernière.

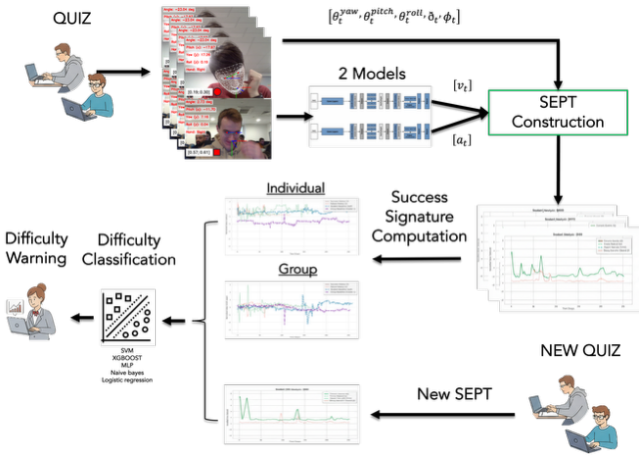


Figure 1 : Prédire la difficulté de l'étudiant pendant l'évaluation en utilisant le cadre SEPT

3.1 Formalisation de l'Espace d'Engage-ment Multimodal

Nous définissons l'engagement multimodal de l'apprenant au temps t comme un vecteur $x_t \in \mathbb{R}^7$ intégrant des états affectifs internes et des marqueurs comportementaux externes dérivés des captures de la webcam. Nous combinons les états internes (l'affect) avec leurs manifestations physiques externes pour capturer l'expérience vécu par l'étudiant :

$$x_t = [v_t, a_t, \theta_t^{yaw}, \theta_t^{pitch}, \theta_t^{roll}, \eth_t, \phi_t]$$

Inférence Affective (Valence v_t/Arousal a_t) : Afin de capturer les dynamiques émotionnelles continues, nous avons développé une architecture de réseau de neurones convolutifs (CNN) spécifiquement optimisée pour l'inférence en temps réel à partir de flux vidéo non contraints (webcam). Face à l'absence de modèles open source, nous avons conçu et entraîné notre propre réseau profond (de bout en bout). Nous nous affranchissons ainsi de la dépendance à des réseaux pré-entraînés génériques souvent sous-optimaux pour cette tâche spécifique. Le modèle repose sur une

architecture efficiente structurée autour de cinq blocs convolutionnels séquentiels.

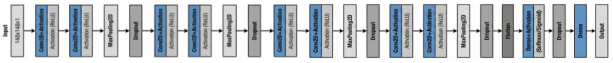


Figure 2 : Architecture du Réseau de Neurones Convolutifs pour la prédiction de la Valence et de l'Arousal

Le modèle a été entraîné sur les jeux de données de référence *AffectNet* [22] et *AffWild* [23]. Il utilise l'erreur quadratique moyenne (MSE) comme fonction de perte et le coefficient de corrélation de concordance (CCC) comme métrique. Comme illustré dans le Tableau 1, nos deux modèles (un pour la valence et un pour l'arousal) atteignent un CCC de 0,714 pour la prédiction de la valence et de 0,644 pour la prédiction de l'arousal. Ces résultats rivalisent avec la performance de l'état de l'art [24], tout en étant spécifiquement calibrés pour la variabilité inhérente aux environnements éducatifs non contrôlés.

Tableau 1. Comparaison de performance des Modèles pour la détection de valence et d'arousal sur *AffectNet* (la meilleure performance est en gras, la seconde soulignée).

	Valence		Arousal	
Modèles	CCC↑	RMSE↓	CCC↑	RMSE↓
CAGE 2024[24]	**0.716**	**0.331**	0.642	**0.305**
VGG-F 2021[22]	0.710	0.356	0.629	0.326
Nos modèles	*0.714*	*0.339*	**0.644**	*0.323*

Caractéristiques de pose de la tête :

- $\theta_t^{yaw}, \theta_t^{pitch}, \theta_t^{roll}$ correspondent aux angles d'Euler de la pose de la tête, indicateurs de l'attention visuelle et de la cognition incarnée (par ex., réflexion, confusion).
- $ð_t$ est la distance normalisée à l'écran, indiquant l'engagement postural (approche/évitement). Nous n'estimons pas la distance absolue visage-caméra de l'étudiant ; au lieu de cela, nous quantifions les changements de distance à partir de l'échelle de la boîte englobant le visage (détecté en utilisant *OPENCV* [25] et *YoloV5* [26]). Sous le modèle perspectif, la taille apparente de l'image d'un objet varie inversement avec la profondeur, donc la taille de la boîte englobante fournit un proxy pratique pour la variation de distance.
- ϕ_t représente l'estimation de l'angle du regard.

Par conséquent, l'engagement de l'étudiant est modélisé comme SEPT [27] $\mathcal{T}$, défini comme une série temporelle ordonnée de longueur L :

$$\mathcal{T}_i = \{x_1, x_2, \dots, x_L\} \in \mathbb{R}^{7\times L}$$

3.2 Modélisation des Signatures de parcours de réussite : Signatures de Groupe vs. Individuelles

Plutôt que d'entraîner un classificateur binaire sur la classe hétérogène "Échec", nous définissons la difficulté comme une déviation d'une "Signature de Réussite". En adoptant ce point de vue, la difficulté académique est une divergence structurelle par rapport à ces signatures de parcours de réussite établies. Nous définissons une « signature de parcours de réussite » construite à partir des données des étudiants les plus performants. Cette approche offre un avantage décisif pour les contextes éducatifs « *Small Data* » : elle ne nécessite des données denses que pour la classe majoritaire (succès), rendant le modèle robuste à la rareté des échantillons d'échec. Nous formalisons deux topologies de référence distinctes : les signatures de parcours de réussite du groupe (M_{group}) et les signatures de parcours de réussite individuelles (μ_{ind}).

Signature de Parcours de Réussite Individuelle (L'Auto-Signature). Nous postulons que chaque étudiant possède sa propre signature d'engagement de réussite. Par exemple, un étudiant peut présenter une agitation productive (mouvements de tête rythmiques pendant la réflexion), tandis qu'un autre reste immobile. Un modèle global pourrait négliger la spécificité individuelle et interpréter l'agitation du premier comme de la distraction. Pour filtrer ces signatures d'engagement de réussite propres, nous construisons une référence basée sur la propre histoire de succès de l'étudiant. Soit ψ_{ind} l'ensemble des trajectoires historiques pour l'étudiant i associées à une performance réussie (note g ≥ 50% de la note maximale). Puisque la moyenne arithmétique de séries temporelles multivariées temporellement déformées est mal définie, nous définissons la Signature de parcours de réussite individuelle μ_{ind} comme le médoïde de l'ensemble ψ_{ind} des trajectoires de l'étudiant. Formellement :

$$\mu_{ind} = arg\,min_{\mathcal{T} \in \psi_{ind}} \sum_{\mathcal{T}' \in \psi_{ind}} D_{DTW}(T, T')$$

μ_{ind} représente la trajectoire historique spécifique qui est la plus centrale au comportement de réussite de l'étudiant, capturant ses habitudes posturales et émotionnelles en l'absence de difficultés académiques.

Signatures de parcours de réussite de Groupe (La Signature Sociale). Pour adresser le problème de démarrage à froid (étudiants sans historique) et pour capturer des modes universels d'engagement, nous définissons une topologie de groupe. Cependant, nous reconnaissons qu'il n'y a pas une seule « bonne façon » d'apprendre ; le succès est multimodal (par ex., « concentration stoïque » vs. « Exploration dynamique »). Nous agrégeons les trajectoires de réussites de toute la cohorte dans un ensemble $\psi_{all} = \cup_i \psi_{ind}$. Nous appliquons un clustering hiérarchique agglomératif sur la matrice de dissimilarité DTW de ψ_{all} pour partitionner l'espace en K clusters distincts $\{C_1, C_2, \dots, C_k\}$. Pour chaque cluster k, nous identifions son médoïde μ_k .

L'ensemble des Signatures de parcours de réussite de groupe est défini comme la collection de ces trajectoires :

$$M_{group} = \{\mu_1, \mu_2, \dots, \mu_k\}$$

Cet ensemble M_{group} forme une variété représentant les signatures de parcours de réussite de groupe distinctes. Celles-ci ne représentent pas une seule « bonne façon » d'apprendre, mais plutôt les modes distincts d'engagement efficace permis au sein du système éducatif. La note g sert uniquement de filtre transitoire pour peupler ψ_{ind} et n'est jamais stockée. Puisque chaque étudiant quel que soit sa moyenne générale présente des séquences réussies, cette approche évite un profilage de capacité ; elle modélise l'alignement momentané de l'engagement, pas le niveau académique général.

3.3 Quantification de la Difficulté comme Déviation

Nous définissons la difficulté comme une déviation par rapport aux signatures de parcours de réussite établies dans la Section 3.2. Le score de difficulté $\delta(T_{new})$ quantifie la dissimilarité entre une nouvelle trajectoire T_{new} et les signatures de référence en utilisant la déformation temporelle dynamique (DTW).

Déviation de Signature Individuelle (Figure 3). Sous l'hypothèse de l'auto-référence, la difficulté est définie comme la déviation par rapport à la signature historique spécifique de l'étudiant μ_{ind}, filtrant efficacement le comportement personnel :

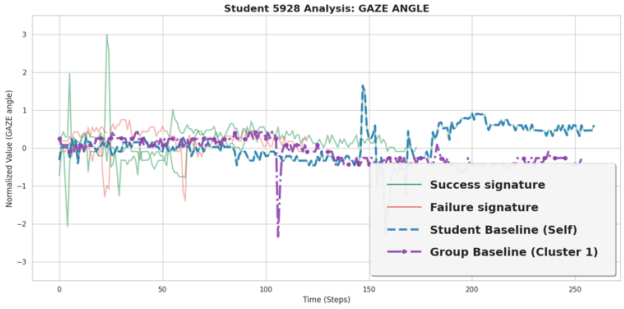
$$\delta_{ind}(T_{new}) = D_{DTW}(T_{new}, \mu_{ind})$$

Cette formulation réduit l'analyse de trajectoire en haute dimension à un scalaire univarié, qui sert d'entrée pour la fonction de décision f décrite dans la section suivante.

Déviation de Signature de Groupe (Figure 3). Sous l'hypothèse de la norme sociale, la difficulté est définie comme la distance à la plus proche Signature de Parcours de Réussite valide au sein de la variété de groupe M_{group}. Cette formulation accommode la nature multimodale du succès :

$$\delta_{group}(T_{new}) = min_{\mu_k \epsilon M_{group}}(D_{DTW}(T_{new}, \mu_k))$$

La déviation de signature est calculée pour chacune des sept dimensions de l'engagement.



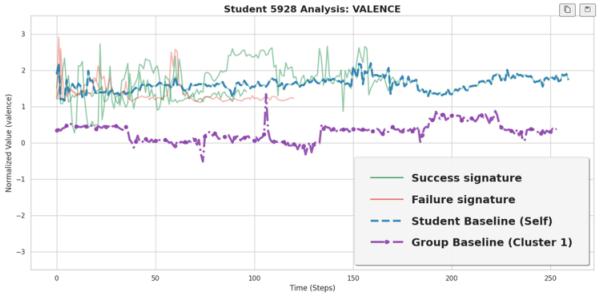


Figure 3 : Déviation de signature de l'Étudiant 5928 (focus sur l'angle du regard et la valence). Les trajectoires d'échec (rouge) montrent une déviation significative de la signature personnelle de l'étudiant (bleu).

Afin d'inférer l'état de difficulté de l'apprenant, le vecteur multivarié des déviations calculé sur les sept dimensions de l'engagement, sert d'entrée à un algorithme de classification supervisée. La valeur cible de cet apprentissage est une variable binaire issue de la vérité terrain objective : la classe positive (y=1 justifiant une alerte) caractérise une trajectoire menant à un échec (score inférieur à 50% de la note maximale), tandis que la classe négative (y=0) désigne une réussite.

Afin d'identifier l'architecture la plus performante pour cette tâche, nous avons évalué et comparé plusieurs familles d'algorithmes, englobant à la fois des modèles linéaires et non linéaires. Plus précisément, nos expérimentations ont porté sur la Régression Logistique et les Machines à Vecteurs de Support (SVM) pour les approches linéaires, ainsi que sur la classification bayésienne naïve, les Forêts Aléatoires, le Perceptron Multicouche (MLP) et l'algorithme XGBoost pour les approches non linéaires.

Dans le cas des modèles linéaire la décision finale du modèle, qui déclenche l'alerte précoce à destination du formateur, repose sur le franchissement d'un seuil de décision définie par un seuil λ. La contribution majeure de notre approche réside dans l'instanciation de ce seuil :

- **Pour l'approche de groupe** : L'algorithme optimise un hyperplan séparateur global et un seuil unique λ_{group} appliqués uniformément à l'ensemble de la cohorte.
- **Pour l'approche individuelle :** L'algorithme intègre une frontière de décision adaptative reposant sur un seuil personnalisé λ_i propre à chaque étudiant *i*. Cette calibration individualisée est cruciale : elle permet de moduler la sensibilité de l'alerte en fonction de la variance comportementale historique inhérente à l'étudiant. Ceci évite les faux positifs pour des apprenants présentant naturellement une forte expressivité physique ou émotionnelle en situation de réussite.

3.4 Stratégie d'Alerte Précoce et Protocole

Pour opérationnaliser la détection de difficultés pour un système de tutorat intelligent (STI) réactif, nous définissons la fonction de prédiction d'échec $f : T \rightarrow \{0,1\}$. Cette

fonction mappe la trajectoire d'un étudiant T_{new} vers un état binaire (0 : Succès/Normal, 1 : Difficulté/Anomalie). Le lien entre la difficulté quantifiée $\delta(T_{new})$ (définie dans la section 3.3 comme la distance aux signatures de référence) et la prédiction f est établi via un seuil de décision λ. Nous postulons qu'un étudiant a des problèmes si sa déviation comportementale excède cette frontière apprise :
$f(T_{new}) = 1[\delta(T_{new}) > \lambda]$

Le choix de cette approche par mesure de similarité élastique (DTW) plutôt que par apprentissage profond (Deep Learning) est motivé par les contraintes spécifiques du contexte pédagogique "*in-the-wild*". Le Tableau 2 positionne SEPT par rapport aux paradigmes dominants de l'état de l'art.

Tableau 2. Positionnement de SEPT face aux études récentes de détection précoce.

Méthode	Caractéristiques techniques	Volume de données requis
Deep Learning (Bi-LSTM, Transformers)[28]	Modélisation de séquences complexes via réseaux profonds.	Très élevé ("Data Hunger").
Ensemble Learning (XGBoost, Random Forest)[29]	Classification basée sur des moyennes statistiques	Moyen.
Notre approche SEPT (OE-DTW)	Alignement élastique et comparaison de signatures.	**Faible ("Small Data").**

Pour valider la robustesse de cette approche, nous établissons deux protocoles expérimentaux dépendant de la complétude de l'observation T_{new} :

- **Analyse Post-Hoc (basée sur la trajectoire complète) :** Le modèle a accès à la trajectoire complète T_{full} de longueur L. Ici, le calcul utilise l'alignement de groupe standard. Cela établit la borne supérieure théorique de performance.
- **Alerte Précoce :** Le modèle observe seulement un préfixe tronqué $T_{partial}$ représentant les premiers 30 % de l'activité.

Dans le scénario d'alerte précoce, appliquer la DTW standard est mathématiquement défectueux. En effet, cette dernière impose une contrainte, obligeant la fin de l'observation partielle (30 % de la chronologie) à s'aligner avec la fin de la signature de référence (100 % de la chronologie). Cela induit des distorsions de déformation temporelle artificielles. Pour atténuer cela, nous employons la DTW à fin ouverte (Open-End DTW ou OE-DTW) [14]. Cette variante relâche la contrainte de point final sur la séquence de référence. Formellement, si Q est la requête étudiante partielle et R est une signature de référence, OE-DTW cherche la longueur de préfixe optimale l au sein de R qui minimise le coût d'alignement :

$$D_{OE-DTW}(Q,R) = min_{l\in[1,|R|]} D_{DTW}(Q, R_{1:l})$$

Cela permet au système de déterminer si le début du comportement de l'étudiant est cohérent avec le début d'une signature de réussite.

3.5 Éthique et Confidentialité : Une Approche « *Privacy-by-Design* »

Le déploiement de systèmes de vision par ordinateur dans les cadres éducatifs soulève des préoccupations légitimes concernant la vie privée et la conformité réglementaire (RGPD en Europe). Pour relever ces défis, cette étude adopte une architecture stricte de *Privacy-by-Design*.
Contrairement aux approches de surveillance traditionnelles basées sur l'enregistrement de flux vidéo bruts, notre pipeline de traitement effectue une extraction immédiate des caractéristiques. Le système ne conserve aucune image du visage de l'étudiant. Seuls des descripteurs numériques anonymisés (points clés du squelette facial - 64 repères 2D, valeurs de valence et d'arousal) sont extraits en temps réel. Cette méthode de minimisation des données assure l'anonymisation de l'identité visuelle de l'étudiant à partir des trajectoires stockées (vecteurs numériques). De plus, l'analyse se limite à la détection de comportement (pose de la tête) et n'opère pas de reconnaissance faciale (identification biométrique), excluant toute utilisation de bases de données d'identité. Cette approche garantit que l'analyse pédagogique bénéficie de la richesse des signaux non verbaux sans compromettre l'anonymat et le droit à l'image des participants, conformément aux directives actuelles sur l'IA de confiance.

4 Résultats Expérimentaux

4.1 Protocole

L'évaluation du cadre SEPT repose sur un corpus de 1849 trajectoires, capturant la dynamique comportementale d'une cohorte d'étudiants en cycle d'ingénierie. Afin de garantir une stricte validité écologique et de simuler les conditions réelles de déploiement d'un système d'alerte précoce, nous avons opté pour un partitionnement temporel des données. Ainsi, l'apprentissage du modèle s'appuie sur un historique constitué de l'intégralité des trois premières évaluations (Quiz 1 à 3), formant un ensemble d'entraînement de 1 279 trajectoires (dont 200 échecs). La capacité de généralisation temporelle et inter-tâches du modèle est ensuite évaluée sur un ensemble de test strictement ultérieur composé des tentatives du dernier quiz (Quiz 4), totalisant 570 trajectoires (dont 80 échecs).

La prévalence de la classe d'intérêt (l'échec) n'étant que de 14,0 % dans l'ensemble de test, ce fort déséquilibre de classe invalide l'utilisation de l'exactitude (accuracy) comme métrique d'évaluation légitime ; en effet, un classificateur majoritaire naïf atteindrait mécaniquement une exactitude de 86 %. Par conséquent, notre protocole d'évaluation s'appuie sur le F1-Score ciblant spécifiquement la classe minoritaire, afin d'évaluer le compromis précision-rappel, ainsi que sur l'Aire Sous la Courbe ROC (AUC) pour quantifier la capacité de discrimination globale et la robustesse du modèle indépendamment de son seuil opérationnel.

Définition de la Trajectoire et Vérité Terrain. Dans le cadre de cette étude, l'unité d'analyse temporelle, désignée comme une « trajectoire », correspond à l'ensemble continu des vecteurs d'état $x_t \in \mathbb{R}^7$ capturés spécifiquement durant le temps de réflexion et de saisie de la réponse pour une seule question. La vérité terrain caractérisant la difficulté a été établie objectivement à l'échelle de l'item cognitif : une trajectoire est étiquetée comme « Échec » (situation de difficulté) si la note obtenue par l'étudiant à la question courante est inférieure à 50 % de la note maximale possible pour cet item. À l'inverse, un score $\geq 50\%$ caractérise une trajectoire de réussite, venant alimenter l'historique de succès individuel ψ_{ind}. Après suppression des séquences inutilisables (occultations du visage, pertes de signal), l'ensemble de données exploitable s'élève à 1849 trajectoires distinctes.

Acquisition et Architecture *Privacy-by-Design*. La capture des flux a été réalisée via un outil de télésurveillance (proctoring) configuré avec un taux d'échantillonnage volontairement bas d'une image par seconde (1 FPS). Cette faible fréquence répond à un double impératif : limiter le bruit à haute fréquence des micro-mouvements non sémantiques. Conformément à notre architecture *Privacy-by-Design*, aucune vidéo brute n'a été stockée. L'inférence des points clés squelettiques et des états affectifs s'est effectuée à la volée, ne persistant que les descripteurs numériques anonymisés.

Conformité Éthique. Le déploiement de ce système a suivi un protocole éthique strict. L'étude a fait l'objet d'une instruction et d'une validation préalable par un comité d'éthique indépendant ainsi que par le Délégué à la Protection des Données de l'établissement. Préalablement à l'expérimentation, l'ensemble des 79 étudiants ont été informés des modalités techniques (absence d'enregistrement vidéo visuel) et ont signés un formulaire de consentement éclairé, en accord avec les directives du RGPD et du futur AI Act.

4.2 Analyse Comparative : Signature de l'étudiant vs. Signature de groupe

Notre hypothèse centrale postule que la difficulté académique se manifeste comme une déviation par rapport aux propres habitudes comportementales d'un étudiant plutôt que comme une déviation par rapport à une signature de groupe. Les résultats valident les Signatures de parcours de réussite Individuelles :

- **Gain de performance :** Le modèle « Étudiant » (régression logistique) atteint un F1-Score de 0,27, surpassant l'approche « groupe » (0,23). Ce gain relatif de +17,4 % confirme que l'historique comportemental personnel est un prédicteur plus granulaire que les moyennes de classe.
- **Pouvoir discriminant :** L'analyse ROC corrobore cette supériorité : alors que l'approche des signatures de parcours de réussite de groupe plafonne à une AUC de 0,61 (SVM), l'approche de signature étudiante monte à 0,67 (SVM). Ce différentiel est critique, car il indique une réduction significative des faux positifs pour le même taux de détection.
- **Stabilité du Modèle :** Les modèles linéaires (SVM, Régression Logistique) surpassent constamment les architectures non linéaires telles que XGBoost ou MLP sur les données de trajectoire basées sur la DTW. Cela suggère que la distance à une « signature personnelle » est une caractéristique linéairement séparable dans le cas des modèles linéaire.

Cette supériorité est corroborée par l'analyse de la courbe ROC présentée en Figure. 4.

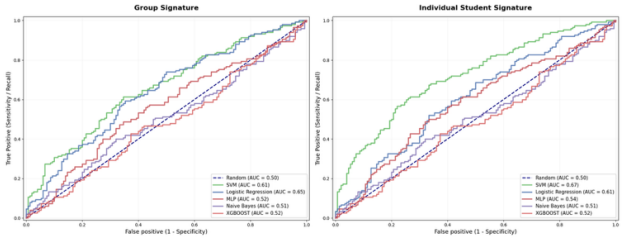


Figure 4 : Courbes ROC. À droite (approche signature Étudiant Individuel), le SVM démontre un pouvoir discriminant significativement meilleur (AUC=0.67). Ce différentiel de +0.06 points d'AUC est critique : il indique que le modèle génère significativement moins de faux positifs pour le même taux de détection.

4.3 Positionnement par rapport aux Études Récentes

Bien que les valeurs absolues (AUC=0.67) puissent paraître modestes, elles doivent être contextualisées par la difficulté intrinsèque de la détection « in-the-wild » (en conditions réelles) utilisant des données conformes à la confidentialité.

Les méthodes utilisant des données démographiques ou des journaux LMS complets atteignent une AUC légèrement plus élevée (0,69-0,71) comme montré dans le Tableau 3. Nous atteignons une performance comparable (seulement 0,04 de différence d'AUC) tout en passant à un paradigme entièrement respectueux de la vie privée. Contrairement à [31], qui dépend du profilage démographique pour le pouvoir prédictif, notre cadre élimine le besoin de données d'attributs personnels.

Tableau 3. Comparaison de SEPT avec des études récentes.

Study	Type de données	Performance (AUC) ↑	Préservation de la confidentialité
SEPT (Our Approach)	Trajectoire émotionnelle et physiques	0.67	**+ +**
Early Warning XGBoost [30]	Socio-Demographique + note du lycée	0.69	- -

MOOC Stacked Ensemble [31]	Journaux complets (LMS) + Données d'information des étudiants	**0.71**	+
EfficientNet Engagement[32]	Vidéo brute (Reconnaissance faciale)	~0.62 (Acc)	- -

4.4 Efficacité de la Détection Précoce (RO2)

Pour qu'un système d'alerte soit utile, il doit identifier les difficultés pendant l'activité. Nous avons testé les modèles sur les premiers 30 % de la trajectoire temporelle.

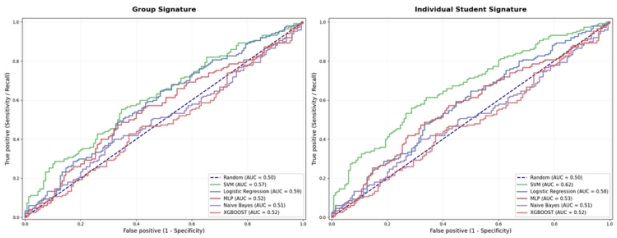


Figure 5 : Performance de détection précoce ($t < 30\%$). La performance de l'approche individuelle reste stable, indiquant que les signaux d'échec sont présents dès le tout début de la tâche.

4.5 Impact Opérationnel : Optimiser les Interventions Ciblées

Au-delà des métriques techniques, la valeur de SEPT réside dans l'optimisation de la « bande passante attentionnelle » limitée d'un instructeur. Dans une simulation d'une cohorte de 70 étudiants avec une prévalence d'échec de 15 % (env. 10 étudiants en difficulté). Pour évaluer cet impact opérationnel, nous analysons la Précision au Rang P (précision *p*) au sein d'un scénario de simulation réaliste. Si un instructeur a la capacité d'effectuer seulement 10 interventions ciblées (p =10), l'efficacité de la stratégie de sélection devient critique.

- Sélection Aléatoire : Un instructeur identifierait seulement 1 ou 2 étudiants en difficulté par hasard.
- Sélection Ciblée SEPT (utilisant les signatures de parcours de réussite Individuelles) : En signalant le Top 10 des étudiants à risque, SEPT identifie correctement 5 étudiants en véritable difficulté.

Cela représente une élévation de performance de 2,5x comparée à une aide humaine non assistée. Dans un cadre « humain dans la boucle » (*human-in-the-loop*), les faux positifs générés par le système ne sont pas des « erreurs » mais des opportunités pour des vérifications bienveillantes avec des étudiants engagés dans une lutte productive. SEPT transforme ainsi une surveillance de masse impossible en un ciblage pédagogique précis et actionnable.

5 Conclusion

Cette étude visait à concevoir et évaluer SEPT, un cadre d'analyse basé sur la trajectoire et intrinsèquement respectueux de la vie privée, pour la détection précoce des difficultés académiques. Au-delà des métriques de performance pures, nos résultats remettent en question l'approche dominante en analyse de données d'apprentissage (*Learning Analytics*), qui tend à évaluer l'engagement étudiant au travers du prisme réducteur d'une norme comportementale globale.

La contribution fondamentale de ce travail réside dans la validation empirique de l'« hypothèse de la signature individuelle ». Nos expériences démontrent que le prédicteur le plus robuste de l'échec académique n'est pas la déviation par rapport à une moyenne de classe, mais la divergence structurelle par rapport à l'historique de succès de l'étudiant lui-même (élévation de performance relative de +17,4 %). Ce résultat s'aligne fortement avec les postulats de la cognition incarnée : l'expression physique de la charge cognitive (ex : faire les cents pas, inclinaisons rythmiques de la tête) est hautement idiosyncratique. Ce constat appelle à un changement de paradigme dans la conception des Systèmes de Tutorat Intelligents (STI) : l'abandon de modèles de classification « *one-size-fits-all* » au profit d'architectures d'auto-étalonnage (*self-benchmarking*), capables de distinguer une "frustration productive" personnelle d'un réel décrochage cognitif.

Nous avons confronté le cadre SEPT aux modèles invasifs de l'état de l'art s'appuyant sur des journaux d'apprentissage complets ou des données socio-démographiques (Tableau 2). Si nous observons un léger compromis prédictif (AUC de 0,67 contre 0,71 pour les méthodes intrusives), nous soutenons que cette différence mathématique marginale masque un avantage opérationnel décisif. Avec l'entrée en vigueur de législations telles que l'AI Act , qui classifie les systèmes d'IA de surveillance éducative et d'évaluation comme présentant un « Haut Risque », les modèles basés sur la reconnaissance faciale biométrique ou le profilage invasif deviennent obsolètes de facto, car légalement et éthiquement impossible à déployer. En s'appuyant exclusivement sur l'extraction éphémère de descripteurs cinématiques et affectifs à basse fréquence (1 FPS), SEPT démontre qu'il est possible de conserver 94 % du pouvoir prédictif de l'état de l'art tout en garantissant un anonymat total des données stockées.

6 Discussion et Travaux Futurs

6.1 Utilité Opérationnelle : L'IA comme aide pédagogique.

La valeur pratique du système a été confirmée dans notre scénario d'Alerte Précoce. En utilisant la méthode de déformation temporelle dynamique ouverte (OE-DTW), SEPT maintient une capacité de discrimination robuste dès les premiers 30 % de l'activité étudiante. En situation de

classe, cela se traduit par une multiplication par 2,5 de la capacité de l'enseignant à cibler les étudiants nécessitant un soutien immédiat. L'objectif de SEPT n'est pas l'automatisation de la notation, mais l'augmentation de la bande passante attentionnelle de l'instructeur. Dans ce cadre « Human-in-the-Loop », les faux positifs générés par le système perdent leur statut d'erreur critique pour devenir de simples opportunités d'interaction pédagogique proactive. L'IA quitte ainsi le rôle de surveillant punitif pour endosser celui d'assistant au diagnostic de la détresse cognitive.

6.2 Limites et Travaux Futurs

Bien que prometteurs, ces résultats comportent plusieurs limites qui orienteront nos travaux futurs.

- **Profil de la cohorte** : Le jeu de données actuel a été collecté auprès d'étudiants en ingénierie face à des tâches convergentes (QCM). Des études ultérieures devront valider la robustesse des trajectoires affectives lors de tâches divergentes ou créatives, notamment en sciences humaines

- **Problème du démarrage à froid (*Cold Start*) :** L'approche par signature individuelle requiert un historique de réussites. Bien que notre système pallie cette lacune transitoire en s'adossant aux "signatures de groupe" lors des premières sessions, la rapidité de convergence vers une signature individuelle stable (nombre minimum de trajectoires requises) reste à modéliser mathématiquement.

- **Résolution temporelle :** L'échantillonnage volontairement parcimonieux à 1 image par seconde, dicté par des contraintes de frugalité et de réduction de l'invasivité, nous prive de l'analyse des micro-expressions faciales (durée < 500 ms). Les travaux futurs pourraient explorer des mécanismes d'attention dynamiques adaptant la fréquence de capture uniquement lors de la détection de transitions affectives critiques.

Divulgation des intérêts. Les auteurs n'ont aucun intérêt concurrent à déclarer qui soit pertinent pour le contenu de cet article.

Références

1. General Data Protection Regulation (GDPR) – Official Legal Text, https://gdpr-info.eu/, last accessed 2024/03/19.
2. Résumé de haut niveau de la loi sur l'intelligence artificielle de l'UE, https://artificialintelligenceact.eu/fr/high-level-summary/, last accessed 2026/01/31.
3. Pekrun, R., Perry, R.P.: Control-value theory of achievement emotions. In: International handbook of emotions in education. pp. 120–141. Routledge/Taylor & Francis Group, New York, NY, US (2014).
4. Camacho-Morles, J., Slemp, G.R., Pekrun, R., Loderer, K., Hou, H., Oades, L.G.: Activity Achievement Emotions and Academic Performance: A Meta-analysis. Educ Psychol Rev. 33, 1051–1095 (2021).
5. Russell, J.: A Circumplex Model of Affect. Journal of Personality and Social Psychology. 39, 1161–1178 (1980).
6. Blumenfeld, P., Paris, A.: School Engagement: Potential of the Concept, State of the Evidence. Review of Educational Research. 74, 59–109 (2004).
7. Wilson, M.: Six views of embodied cognition. Psychonomic Bulletin & Review. 9, 625–636 (2002).
8. Zou, L., Zhang, Z., Mavilidi, M., Chen, Y., Herold, F., Ouwehand, K., Paas, F.: The synergy of embodied cognition and cognitive load theory for optimized learning. Natural Human Behavior. 9, 877–885 (2025).
9. Toward Semantic Gaze Target Detection, https://www.proceedings.com/079017-3858.html, last accessed 2026/01/30.
10. Holstein, K., Aleven, V., Rummel, N.: A Conceptual Framework for Human–AI Hybrid Adaptivity in Education. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., and Millán, E. (eds.) Artificial Intelligence in Education. pp. 240–254. Springer International Publishing, Cham (2020).
11. Zaletelj, J., Košir, A.: Predicting students' attention in the classroom from Kinect facial and body features. Journal of Image Video Processing. 2017, 80 (2017).
12. Mohamad Nezami, O., Dras, M., Hamey, L., Richards, D., Wan, S., Paris, C.: Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression. Proc. Of ECML PKDD. pp 273-289 (2020).
13. Akpanoko, C.E., S., A.T., Cordell, G., Biswas, G.: Investigating the Relations between Students' Affective States and the Coherence in their Activities in Open-Ended Learning Environments. Proc of the 17th International Conference on Educational Data Mining (2024).
14. Grafsgaard, J., Wiggins, J., Boyer, K., Wiebe, E., Lester, J.: Automatically Recognizing Facial Indicators of Frustration: A Learning-Centric Analysis. (2013).
15. Basavaiah, J., Anthony, A.A., N, N.K.H., Mahadevaswamy, Patil, C.M.: Facial Emotion Recognition: A Review on State-of-the-art Techniques. In: 2024 4th International Conference on Data Engineering and Communication Systems (ICDECS). pp. 1–6 (2024).
16. Nadaud, E., Yaacoub, A., Haidar, S., Grand, B., Prevost, L.: Emotion Trajectory and Student Performance in Engineering Education: A Preliminary Study. Int. Conf. on Research Challenges in Information Science, 410-424 (2024).
17. Zambrano, A.F., Ocumpaugh, J., Baker, R.S., Vandenberg, J.: Better to Be Confused or Frustrated Than

Bored: Analyzing Affect Dynamics Across Player Archetypes. In: Carmona, G., Lima, C., Santos, M.J., Benítez, H., Montero-Moguel, L., and Galarza-Tohen, B. (eds.) Advances in Quantitative Ethnography. pp. 384–399. Springer Nature Switzerland, Cham (2026).
18. Baker, R.S.J. d., D'Mello, S.K., Rodrigo, Ma.M.T., Graesser, A.C.: Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. International Journal of Human-Computer Studies. 68, 223–241 (2010).
19. Piot, M., Alabarbe, T., Gonzalez, J., le Bail, C., Prevost, L., Bourdeau, J., Bernard, F.X., Baker, M., Detienne, F.: Joint analysis of verbal and nonverbal interactions in collaborative E-learning. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). pp. 1–5 (2019).
20. D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. Learning and Instruction - LEARN INSTR. 22, (2012).
21. Zhao, J., Itti, L.: shapeDTW: shape Dynamic Time Warping, http://arxiv.org/abs/1606.01601, (2016).
22. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Trans. Affective Comput. 10, 18–31 (2019).
23. Liu, M., Kollias, D.: Aff-Wild Database and AffWildNet, http://arxiv.org/abs/1910.05318, (2019).
24. Wagner, N., Mätzler, F., Vossberg, S.R., Schneider, H., Pavlitska, S., Zöllner, J.M.: CAGE: Circumplex Affect Guided Expression Inference, http://arxiv.org/abs/2404.14975, (2024).
25. OpenCV: Cascade Classifier, https://docs.opencv.org/3.4/db/d28/tutorial_cascade_classifier.html, last accessed 2025/03/26.
26. What is YOLOv5: A deep look into the internal features of the popular object detector, https://arxiv.org/html/2407.20892v1, last accessed 2025/03/26.
27. Nadaud, E., Yaacoub, A., Legrand, B., Prevost, L.: SEPT: Uncovering Student Difficulties through Emotional and Physical Trajectories during Online Assessments. (2025).
28. Kang, X., Nie, Y.: Design and analysis of teaching early warning system based on multimodal data in an intelligent learning environment. PeerJ Comput. Sci. 11, e2692 (2025).
29. Chang, Y.-H., Chen, F.-C., Lee, C.-I.: Developing an Early Warning System with Personalized Interventions to Enhance Academic Outcomes for At-Risk Students in Taiwanese Higher Education. Education Sciences. 15, (2025).
30. Carballo-Mendívil, B., Arellano-González, A., Ríos-Vázquez, N.J., Lizardi-Duarte, M. del P.: Predicting Student Dropout from Day One: XGBoost-Based Early Warning System Using Pre-Enrollment Data. Applied Sciences. 15, 9202 (2025).
31. Ajayi, O.O.: Predicting Student Dropout Risk in Online Learning using Stacked Ensemble Machine Learning and Explainable AI Techniques. International Journal of Computer Applications. 187, 26–29 (2025).
32. Alyuz, N., Okur, E., Genc, U., Aslan, S., Tanriover, C., Esme, A.A.: Unobtrusive and Multimodal Approach for Behavioral Engagement Detection of Students, http://arxiv.org/abs/1901.05835, (2019).

La troisième voie des « Code LLM » ouverts et éthiques : une stratégie de différenciation durable face aux modèles extractivistes ?

Robert Viseur[1]

[1] UMONS, Faculté Warocqué d'Économie et de Gestion

robert.viseur@umons.ac.be

Résumé

L'article examine si la publication de « Code LLM » respectueux des licences libres ou open-sources peut constituer une stratégie viable de différenciation face aux modèles extractivistes. Il montre que la conformité juridique et la prise en compte des communautés (licences, attribution, pratiques de collecte) renforcent la légitimité et l'acceptabilité. Toutefois, leur performance encore inférieure limite aujourd'hui leur capacité à peser sur le marché des assistants à la programmation.

Mots-clés

Code LLM, open model, droit d'auteur, FLOSS, StarCoder.

Abstract

This paper examines whether releasing Code LLMs that respect free and open-source software licences can constitute a viable differentiation strategy vis-à-vis extractivist models. It finds that legal compliance and attention to community norms (licensing, attribution, and data-collection practices) strengthen legitimacy and acceptability. However, their still-lacklustre performance currently limits their ability to make an impact in the programming assistant market.

Keywords

Code LLM, open model, copyright, FLOSS, StarCoder.

1 Introduction

Le terme « *IA générative* » désigne « *des techniques informatiques capables de générer à partir de données d'entraînement des contenus apparemment nouveaux et significatifs* » (p.111) [4]. Trois niveaux sont distingués [4]. Le premier concerne le modèle, par exemple le *Large Langage Model* (LLM) comme GPT-5.5. Le second est relatif au système technique, par exemple l'agent conversationnel tel que ChatGPT, permettant d'interagir avec le modèle. Le troisième est le niveau applicatif relatif aux problèmes résolus avec l'IA générative comme, dans le cas qui nous intéresse davantage, la production d'un code source fonctionnel. La diffusion des modèles d'IA générative s'est accompagnée de nombreux conflits juridiques, qu'il s'agisse de modèles de diffusion ou de modèles de langage [5]. Citons en particulier le litige « *DOE 1 et al v. GitHub, Inc. et al* » portant notamment sur la réutilisation de codes sources sous licences libres pour l'entraînement du modèle CODEX.

Les LLM peuvent donc, comme les Code LLM, être spécialisés. Les « Code LLM » sont notamment capables de transformer en code source des spécifications en langage naturel [7, 8]. Les « Code LLM » ont suscité un intérêt croissant du fait, d'une part, des gains de productivité espérés dans les activités de développement [13], d'autre part, de la démocratisation de la programmation grâce à des approches *low / no code* telles que le *vibe coding* [16]. Parmi les outils les plus connus, citons GitHub Copilot, présenté en juin 2021, puis intégré à la plateforme GitHub, basé sur le « Code LLM » CODEX produit par OpenAI [13]. D'autres ont suivi comme Amazon CodeWhisperer ou Star Coder [7]. Ces modèles sont entraînés sur de vastes corpus de données, notamment composés de codes sources, généralement collectés massivement sur le Web [7]. Dans ce contexte, le modèle Star Coder se distingue par l'attention portée au respect des droits d'auteur et des licences logicielles, tentant ainsi d'échapper à la critique de l'extractivisme des *bigtechs* [2].

Notre communication portera dès lors sur la question de recherche suivante : « *Le choix de publier des « Code LLM » respectueux des licences libres et open-sources constitue-t-il une stratégie viable de différenciation face aux modèles extractivistes ?* ». Notre recherche sera découpée en trois sections. Dans une première section, nous reviendrons sur les bases juridiques liées à la protection des œuvres de l'esprit, en particulier les logiciels, incluant les logiciels libres et open-sources. Dans une seconde section, nous présenterons les griefs portant sur les producteurs de LLM puis, plus spécifiquement de « Code LLM ». Sur cette base, dans une troisième section dédiée à une discussion, et avant de conclure, nous discuterons notre question de recherche au regard des éléments précédents.

2 Propriété intellectuelle

Toute œuvre de l'esprit (un roman, un article scientifique, une photographie, etc.) est automatiquement protégée par droit d'auteur dès lors qu'elle découle d'une activité humaine, a été mise en forme et répond à des exigences minimales d'originalité [9, 12]. L'application du droit d'auteur fait l'objet d'un effort d'homogénéisation à l'aide de traités internationaux comme la Convention de Berne [1], Certains détails diffèrent cependant entre le droit d'auteur étasunien (*copyright*) et le droit d'auteur continental. Le *copyright* reconnaît ainsi le *fair use*, une exception créée par la jurisprudence. Quatre facteurs non

exclusifs sont pris en compte pour déterminer si un usage particulier est équitable [1] : (1) le but et la nature de l'usage (commercial ou non), (2) la nature du matériel protégé, (3) la quantité et l'importance du matériel utilisé et (4) les effets de l'utilisation sur le marché potentiel ou la valeur du travail protégé. Ce dernier critère peut d'ailleurs être rapproché du triple test de la Convention de Berne. Côté continental, cette exception n'existe pas. Cependant, dans le cas particulier des activités de *Text and Data Mining* (TDM), le *scraping* de ressources en ligne est autorisé dans un cadre commercial dès lors que l'*opt out*, mis par exemple en œuvre avec le protocole d'exclusion des robots, est respecté [17].

Les logiciels constituent aussi des œuvres de l'esprit, par défaut protégées par le droit d'auteur. Le caractère libre ou propriétaire d'un logiciel découle de la licence qui encadre les usages autorisés [11]. Les logiciels propriétaires relèvent du régime le plus courant : l'éditeur conserve un contrôle étroit sur l'exploitation du programme afin d'en tirer des revenus via la concession de licences. Ces licences prévoient généralement des restrictions, telles qu'un usage soumis à redevance, l'interdiction de copier, l'absence d'accès au code source, ainsi que l'interdiction de modifier et de redistribuer le logiciel. À l'inverse, les logiciels libres se définissent par les libertés garanties à l'utilisateur : la liberté d'utiliser le logiciel, la liberté d'accéder au code source pour l'étudier ou le modifier, et la liberté de redistribuer des copies, gratuitement ou non (FSF). Certaines licences libres (p. ex. GPL) intègrent un mécanisme de *copyleft* visant à empêcher l'appropriation, en imposant lors de la redistribution la mise à disposition du code source et le maintien de la même licence. À l'inverse, les licences permissives (p. ex. BSD) autorisent plus facilement la réutilisation et l'intégration dans des logiciels propriétaires, sans obligation de réciprocité. Enfin, les licences hybrides combinent des clauses inspirées du libre et du propriétaire : elles peuvent autoriser l'usage, la copie et l'accès au code, tout en limitant certaines modalités de distribution. Elles ne satisfont donc pas pleinement aux critères du libre, sans relever entièrement du propriétaire.

3 Conflits juridiques et éthiques

La diffusion des LLM s'est, d'une manière générale, accompagné de conflits judiciaires autour de l'utilisation de données massivement collectées sur le Web pour l'entraînement des modèles. Les modèles de la famille GPT sont ainsi substantiellement entraînés à partir de données du Common Crawl, une archive du Web collectée par l'association éponyme, mais aussi sur des données scrapées par les producteurs eux-mêmes [17]. En ont très rapidement résulté des procès entre producteurs de modèles et éditeurs à l'image de l'action opposant OpenAI et The New York Times [1]. La plupart des litiges sont actuellement localisés aux USA. Les producteurs y justifient leurs actions par le *fair use* tandis que les auteurs ou éditeurs critiquent cette interprétation, en plus de relever de fréquents problèmes de régurgitation de données d'entraînement, ce qui constitue une violation des droits d'auteur. Côté européen, les producteurs sont protégés par l'exception TDM dès lors que l'*opt-out* est respecté. Les régurgitations excèdent cependant cette exception dès lors qu'elles suggèrent une mémorisation durable des données d'entraînement au sein du modèle.

Dans ce contexte tendu, les éditeurs ont progressivement établi un rapport de force en interdisant l'accès à leurs sites. Pour ce faire, des dispositifs de blocage actif ou passif ont été massivement déployés, à l'image du protocole d'exclusion des robots reconnu comme mécanisme valide d'*opt-out* [17]. Les tensions s'aplanissent cependant progressivement grâce au développement d'un marché de la licence permettant aux producteurs de payer pour disposer d'un accès légal à des données d'entraînement ou de contexte [14].

Dans le cas particulier des logiciels libres, d'autres griefs sont venus s'ajouter. Premièrement, en plus de la polémique sur la violation du droit d'auteur dans les activités de collecte puis d'entraînement, certains responsables de projets libres ont dénoncé les pratiques de collecte agressive (« *agressive crawling* ») conduisant à drainer les ressources des serveurs des projets voire conduisant à l'indisponibilité des dépôts Git autohébergés[1]. Ce phénomène s'explique notamment par des dispositifs tels que Mellowtel, permettant la monétisation des extensions de navigateurs en transformant ces derniers en *crawlers* discrets difficilement détectables. Deuxièmement, certains projets se sont plaints de la croissance de l'« *AI slop* », c'est-à-dire de contenus de faible qualité produits par IA générative. Le projet curl, par exemple, a ainsi vu le nombre de contributions augmenter mais celui des propositions exploitables se réduire, ce qui provoque une dégradation du rapport signal sur bruit dans les contributions au projet, source d'épuisement des mainteneurs [16]. Troisièmement, les capacités d'un « Code LLM » peuvent pénaliser le modèle d'affaires des éditeurs, comme dans le cas de Tailwind CSS (forte popularité mais réduction du chiffre d'affaires de 80 % et licenciement de 75 % des ingénieurs[2]). Ces constats crédibilisent la critique de l'extractivisme des *bigtechs* [2], c'est-à-dire la collecte de données, sans considération pour les ressources liées à leur hébergement ni contribution utile.

Lorsque la licence est mentionnée, son respect peut être exigé. Ainsi une *class action* est menée par des développeurs contre la plateforme GitHub (« *DOE 1 et al v. GitHub, Inc. et al, No. 4:2022cv06823 - Document 195 (N.D. Cal. 2024)* ») ainsi que contre OpenAI, producteur du modèle CODEX, et Microsoft, producteur de Copilot et propriétaire de la plateforme GitHub depuis 2018 [3]. Le premier grief concerne la vente supposée de code source par GitHub malgré les conditions générales d'utilisation (*Term of Use*, ToS) de la plateforme. Le second grief concerne le non-respect de la licence open-source car, d'une part, le producteur réalise une reproduction du code source (entraînement), d'autre part, l'IA générative est parfois sujette aux régurgitations de code source (génération), et ce, sans attribution (paternité) ni mention

[1] Cf. https://www.theregister.com/2025/03/18/ai_crawlers_sourcehut.

[2] Cf. https://www.indiatoday.in/ (...) -project-2849230-2026-01-09.

correcte de la licence [3]. En pratique, seul un des plaignants s'est révélé capable de démontrer qu'un *prompt* conduit à une réponse plagiant son code source tandis que seuls deux autres (sur les cinq) se sont révélés capables de démontrer la similitude [6]. Or, l'application du DMCA suppose une reproduction à l'identique [3, 6]. Finalement, seule la violation de la licence open-source a été conservée par le juge.

Reste que la mention de la licence n'est pas systématique, même sur GitHub. Ainsi, en pratique, l'affectation d'une licence aux logiciels publiés en ligne n'est pas systématique. Par exemple, lors d'une étude réalisée en 2013, seul 14,9 % des répertoires de code source sur GitHub présentait un fichier de licence à la racine du projet [10]. Parmi ceux-ci, près d'un tiers (28 %) mentionnait la licence dans le fichier README plutôt que dans un fichier LICENSE ou COPYING, comme recommandé. La plupart du code publié sur GitHub n'était donc pas open-source. Une étude menée en 2024 sur des plateformes de gestion de *package* a montré que l'absence de licence restait endémique dans certains environnements comme PyPi mais s'améliorait sur d'autres comme RubyGems [19]. Dans le cas des deux études, les licences permissives comme MIT et Apache sont privilégiées par les développeurs. Or, un code publié, en l'absence de licence explicite, est soumis au droit d'auteur et doit donc être considéré comme un code source propriétaire. Vendome confirme la forte diffusion de la licence permissive Apache sur GitHub [15].

4 Vers des « Code LLM » plus éthiques ?

Le respect de la paternité et des licences d'un logiciel est cependant au cœur de certains projets d'intelligence artificielle. C'est notamment le cas du jeu de données The Stack v2 [8]. Mené dans le cadre du BigCode Project, The Stack v2 comprend : (1) du code source provenant de l'archive de Software Heritage, avec un filtrage par licence pour ne garder que les codes sous licence permissive (donc rien sans licence ni *copyleft*) après détection via ScanCode[3] ; (2) des *issues* GitHub reprises de GHArchive ; (3) des *pull requests* de GitHub reprises de GHArchive ; des Notebooks de Jupyter extraits de l'archive de Software Heritage ou publiés par Kaggle ; (4) des *crawls* de documentation, incluant Read The Docs pour sa clarté ; (5) des extraits de StackOverflow et de RedPijama. Cette diversité de données est notamment nécessaire pour pouvoir relier des consignes de codage aux codes appropriés. Ce jeu de données est par exemple utilisé par le modèle StarCoder 2.

L'immixtion de l'IA générative dans les projets de logiciels libres s'est exposée à de fortes résistances comme l'attestent les exemples des projets curl ou Gentoo [16]. L'exposé des principaux griefs a par exemple été synthétisé au sein de l'« *AI Policy* » de Gentoo Linux. Ces griefs concernent, premièrement, la problématique du respect du droit d'auteur des réponses (« *copyright concerns* »), deuxièmement, la médiocrité des codes sources générés (« *quality concerns* ») et, troisièmement, diverses problématiques éthiques (« *ethical concerns* »). Ces dernières incluent à nouveau la problématique du droit d'auteur, cette fois-ci pour le volet de la provenance des données d'entraînement (et plus le manque de traçabilité des contributions en lien avec le risque de régurgitation). Cette préoccupation ressort de nombreuses discussions où la GenAI est perçue comme une sorte de « *copyright laundering machine* » imposée par les *bigtechs*.

Sur le plan de la conformité juridique, un modèle respectueux des droits d'auteurs des contributeurs et des licences (en excluant des données d'entraînement des logiciels couverts par des licences à fort devoir contributif) favorise l'accueil dans le contexte de projets libres puisqu'il répond aux préoccupations légales et éthiques. De même, le soin accordé par BigCode Project aux procédures de collecte respectueuse répond aux préoccupations émergentes liées aux pratiques de collecte agressive. Par contre, ces modèles ne semblent pas aptes, à l'heure actuelle, de répondre au grief relatif à la qualité des réponses. Ainsi, les modèles entraînés sur base de *datasets* plus complets démontrent encore des performances sensiblement supérieures. Jiang et ses co-auteurs mesurent ainsi, pour le test « Pass@1 » (celui-ci représente le pourcentage pour lequel le modèle de langage utilisé une seule fois a pu générer la bonne réponse) un score de 92,0 pour Claude 3.5-Sonet (meilleur modèle propriétaire) contre 90,2 à DeepSeek-Coder-V2-Instruct (meilleur modèle ouvert[4]) mais 72,6, 46,3 et 33,6 pour respectivement StarCoder2-Instruct, StarCoder2 et StarCoder [7]. Le modèle DeepSeek-Coder-V2, bien qu'ouvert, s'appuie sur une collecte de données plus large [20]. Son corpus de pré-entraînement, notamment basé sur GitHub et Common Crawl, comprend ainsi 60 % de code source, 10 % de mathématiques et 30 % de langage naturel. Le filtrage des données collectées est davantage justifié par des critères de qualité que de conformité aux licences sans qu'un audit plus approfondi soit possible, du fait du manque de transparence. L'exigence de conformité semble dès lors moins passer par l'ouverture des modèles que par une meilleure gouvernance des données utilisées pour leur entraînement.

5 Conclusion

Cette étude suggère que la publication de « Code LLM » respectueux des licences ouvre une voie de différenciation principalement fondée sur la légitimité (liée à la conformité, à l'attribution et aux pratiques de collecte) et sur une meilleure acceptabilité auprès des communautés associées aux projets de logiciel libre. Néanmoins, tant que persiste un écart de performance avec les modèles entraînés sur des corpus plus larges et juridiquement ambigus, cette différenciation reste fragile.

[3] Cf. https://scancode-toolkit.readthedocs.io/en/latest/getting-started/home.html. Cette vérification est justifiée du point de vue de l'objectif de conformité étant donné les incohérences dans la documentation juridique des projets sur GitHub [18].

[4] Nous utilisons le terme modèle ouvert (« *open model* ») pour inclure tant les modèle open-sources (au sens de l'Open Source AI Definition) que des modèles open-weights soumis à des licences moins ouvertes comme la licence OpenRAIL couvrant StarCoder2.

La viabilité de cette stratégie dépendra donc de la capacité à réduire cet écart et à inscrire ces modèles dans une chaîne d'outils capable d'articuler garanties de conformité et assistance au codage au sein d'un écosystème de confiance. Dans un contexte de contentieux et d'incertitude réglementaire, cette approche apparaît comme une stratégie de gestion du risque attractive pour les fournisseurs et pour les organisations, en particulier dans les segments de marché sensibles à la conformité (p. ex. secteur public). Cependant, cette différenciation demeure principalement défensive : elle permet de réduire les risques juridiques, réputationnels et d'acceptabilité associés aux modèles extractivistes, sans remettre en cause les mécanismes plus fondamentaux de captation de valeur sur lesquels repose l'économie des « Code LLM ».

6 Références

[1] Al-Busaidi, A. S., Raman, R., ... & Walton, P. (2024). Redefining boundaries in innovation and knowledge domains: Investigating the impact of generative artificial intelligence on copyright and intellectual property rights. Journal of Innovation & Knowledge, 9(4), 100630. https://doi.org/10.1016/j.jik.2024.100630.

[2] Chandrasekhar, R. (2025). Legal frictions for data openness: Reflections from a case-study on re-use of the open web for AI training. Centre Internet et Société - CNRS; Inno3; Open Knowledge Foundation, pp.75. https://dx.doi.org/10.5281/zenodo.15097649.

[3] Farcon, J. F. (2024). Attribution Or Attrition? Doe 1 V. Github, Inc. As A Case For A Robust, Horizontal, Moral Right Of Attribution In Gen AI. https://dx.doi.org/10.2139/ssrn.4946503.

[4] Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. Business & Information Systems Engineering, 66(1), 111-126. https://doi.org/10.1007/s12599-023-00834-7.

[5] Freeman, J., Rippe, C., Debenedetti, E., & Andriushchenko, M. (2024). Exploring memorization and copyright violation in frontier LLMs: A study of the New York Times v. OpenAI 2023 lawsuit. arXiv preprint arXiv:2412.06370. https://doi.org/10.48550/arXiv.2412.06370.

[6] Guildea, B. (2025). The problem of pleadings in AI copyright litigation: lessons from Getty v Stability (UK) and Doe v GitHub (US). Journal of Intellectual Property Law and Practice, jpaf022. https://doi.org/10.1093/jiplp/jpaf022.

[7] Jiang, J., Wang, F., Shen, J., Kim, S., & Kim, S. (2024). A survey on Large Language Models for code generation. arXiv preprint arXiv:2406.00515. https://doi.org/10.48550/arXiv.2406.00515.

[8] Lozhkov, A., Li, R., Allal, L. B., Cassano, F., Lamy-Poirier, J., Tazi, N., ... & de Vries, H. (2024). Starcoder 2 and The Stack v2: The next generation. arXiv preprint arXiv:2402.19173. https://doi.org/10.48550/arXiv.2402.19173.

[9] Mattatia, F. (2017), Droit d'auteur & propriété intellectuelle dans le numérique, Paris, Éditions Eyrolles. ISBN : 978-2-212-67426-2.

[10] McAllister, N. (2013). Study: Most projects on GitHub not open source licensed. The Register, 18 avril 2013. https://www.theregister.com/2013/04/18/github_licensing_study/.

[11] Muselli, L. (2008). Le rôle des licences dans les modèles économiques des éditeurs de logiciels open source. Revue Française de Gestion, 181(1), 199-214. https://doi.org/10.3166/rfg.181.199-214.

[12] Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024). Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. Computer Law & Security Review, 55, 106066. https://doi.org/10.1016/j.clsr.2024.106066.

[13] Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of ai on developer productivity: Evidence from GitHub Copilot. arXiv preprint arXiv:2302.06590. https://doi.org/10.48550/arXiv.2302.06590.

[14] Stratton, M. (2025). Market-Based Licensing for Publishers' Works Is Feasible. Big Tech Agrees. Colum. JL & Arts, 48, 434. https://doi.org/10.52214/jla.v48i4.13925.

[15] Vendome, C. (2015). A large scale study of license usage on GitHub. In 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering (Vol. 2, pp. 772-774). IEEE. https://doi.org/10.1109/ICSE.2015.245.

[16] Viseur, R. (2026). Comment transformer l'« AI slop » en « AI fit » ? Régulation socio-technique des contributions génératives dans les projets libres. Actes de la 31ème conférence de l'AIM, Neuchâtel (Suisse), 20-22 mai 2026.

[17] Viseur, R., & Finet, A. (2025). Dépasser ou respecter la norme? Une analyse de la stratégie de plateformes numériques sous l'angle de la théorie néo-institutionnelle. Innovations, 2025/3 N° 78. https://doi.org/10.3917/inno.pr2.0193.

[18] Wolter, T., Barcomb, A., Riehle, D., & Harutyunyan, N. (2023). Open source license inconsistencies on github. ACM Transactions on Software Engineering and Methodology, 32(5), 1-23. https://doi.org/10.1145/3571852.

[19] Wu, J., Bao, L., Yang, X., Xia, X., & Hu, X. (2024). A large-scale empirical study of open source license usage: Practices and challenges. In Proceedings of the 21st International Conference on Mining Software Repositories (pp. 595-606). https://doi.org/10.1145/3643991.3644900.

[20] Zhu, Q., Guo, D., Shao, Z., Yang, D., Wang, P., Xu, R., ... & Liang, W. (2024). Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. arXiv preprint arXiv:2406.11931. https://doi.org/10.48550/arXiv.2406.11931.

Une Balance à levier comme neurone : Un système mécanique pédagogique pour l'enseignement des principes fondamentaux de l'IA

Thierry Bénard[1]

[1] Fondation CGénial

thierry.b.benard@gmail.com

Résumé

L'enseignement des concepts fondamentaux de l'intelligence artificielle peut être rendu difficile par l'abstraction mathématique des modèles. Nous proposons un système mécanique basé sur une balance à levier qui matérialise physiquement le fonctionnement d'un neurone artificiel. Le dispositif établit une correspondance entre les paramètres d'apprentissage (masse d'entrée, position du contrepoids, angle du levier) et les composants d'un neurone (entrée, poids synaptique, sortie). La simplicité du système permet deux approches pédagogiques : d'une part, une modélisation analytique complète par les équations de couple, et d'autre part, un apprentissage itératif par descente de gradient manuelle. Ce dispositif offre une approche kinesthésique pour enseigner la régression linéaire sans recours à l'informatique, améliorant ainsi l'appropriation des concepts de base de l'IA. Pour des références complémentaires, voir le dépôt GitHub : TorqueBalance-Neuron.

Mots-clés

cognition incarnée, neurone mécanique, descente de gradient, pédagogie de l'IA, activités débranchées.

Abstract

Teaching fundamental concepts of artificial intelligence often faces the challenge of model's mathematical abstraction. We propose a mechanical system based on a lever balance that physically materializes the operating of an artificial neuron. The device establishes a correspondence between learning parameters (input mass, counterweight position, lever angle) and neuron components (input, synaptic weight, output). The system's simplicity enables a dual pedagogical approach : on one hand, complete analytical modeling through torque equations, and on the other hand, iterative learning through manual gradient descent. This device offers a kinesthetic approach for teaching linear and logistic regression without computer use, thereby improving basic AI's concepts. For further references see Github repo

Keywords

AI literacy, embodied cognition, mechanical neuron, gradient descent, CS unplugged.

1 Introduction

Les enjeux sociétaux créés par l'utilisation de l'IA rendent nécessaire une meilleure compréhension, par les citoyens, du fonctionnement de cette technologie. Dans le cadre des démarches visant à promouvoir une IA explicable, des travaux soulignent que l'explicabilité ne peut pas être réduite à la seule transparence algorithmique, mais qu'elle dépend aussi de la compréhension conceptuelle des utilisateurs. Dans cette perspective, l'alphabétisation à l'IA (*AI Literacy*) a été alors proposée comme un complément nécessaire à l'XAI afin que les explications soient significatives et interprétables pour des utilisateurs non experts [9].

Cependant, l'enseignement de ces concepts se heurte à une barrière importante : l'abstraction mathématique des modèles. L'enjeu est donc de permettre aux apprenants de développer une compréhension conceptuelle des systèmes d'apprentissage automatique [2], notamment la "*Big Idea #3 : Computers can learn from data*" définie dans [10] comme l'un des cinq concepts majeures de l'IA.

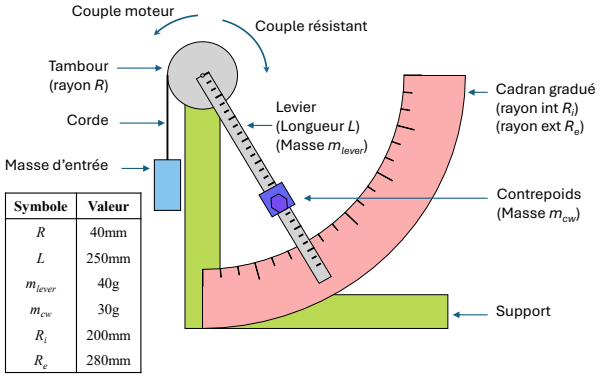
Alors que les approches dites "débranchées" (*unplugged* pour sans ordinateur) se développent pour rendre l'IA accessible sans ordinateur [3], peu de dispositifs matérialisent physiquement les opérations mathématiques sous-jacentes au neurone artificiel. Nous proposons ici un système mécanique basé sur une balance à levier. Conformément aux recommandations internationales appelant à développer des approches pédagogiques adaptées au niveau des apprenants [4], le dispositif offre une voie d'accès aux fondements de l'IA sans recours à l'informatique (*AI unplugged*). Dans la continuité des dispositifs mécaniques tels que la planche de Galton [7], qui matérialise physiquement des lois statistiques ou le réseau de neurone en bois de Schaffland et al. [8] ou encore le réseau de neurone en fil de laine de [5], nous proposons un neurone mécanique basé sur une balance à levier. Le dispositif illustre la capacité d'un système à apprendre à partir des données qui est la propriété essentielle de l'apprentissage automatique. Cette dernière n'étant qu'une des approches de l'IA qui inclue aussi une approche symbolique. Ce dispositif s'inscrit dans les méthodes d'apprentissage concrètes, où la manipulation facilite la compréhension conceptuelle, en particulier en mathé-

matiques [1]. Il prend l'exemple de la régression linéaire pour introduire une procédure d'apprentissage inspirée de la descente de gradient, mais réalisable manuellement par les élèves et sans recours à l'ordinateur. Ce travail s'inscrit dans une démarche de recherche axée sur la conception, qui met l'accent sur la conception itérative et l'évaluation préliminaire d'un support pédagogique. Cette étude pilote présente une première évaluation conduite auprès d'élèves du cycle 4. Cette phase exploratoire vise à identifier des indicateurs pertinents d'appropriation des notions fondamentales de l'IA, sans ambition de généralisation statistique des résultats qui devra faire l'objet d'une étude plus complète.

2 Description du système mécanique

2.1 Architecture Générale

Le système proposé (figure 1 et figure 3) s'inspire des balances à levier coudé de type "weighing scaling bent-lever" [6]. Il s'agit d'un instrument de mesure de la masse exploitant l'équilibre de deux couples mécaniques dont l'un est dépendant de l'angle.
L'ensemble est monté sur un pivot fixe O constituant l'axe de rotation commun au tambour et au levier. Le système est conçu pour fonctionner dans une plage angulaire de 0° à 90°, limitée par des butées mécaniques.



Symbole	Valeur
R	40mm
L	250mm
m_{lever}	40g
m_{cw}	30g
R_i	200mm
R_e	280mm

FIGURE 1 – Dispositif vue de face

2.2 Les composants et leurs caractéristiques

Tambour. Cylindre de rayon R monté coaxialement avec le levier. Le rayon R est dimensionné pour produire une sensibilité suffisante dans un encombrement raisonnable.

Levier. Barre rigide de longueur L fixée au tambour, de masse m uniforme. Sa longueur est munie d'une glissière avec graduations permettant un positionnement mesurable du contrepoids.

Contrepoids. Masse m_{cw} montée sur une fixation à vis de serrage. La position radiale r_{cw} constitue le paramètre d'apprentissage du système. Un système d'indexation sur une échelle graduée facilite les réglages et la reproductibilité des mesures.

Corde et crochet. Corde s'enroulant tangentiellement autour du tambour. Un crochet est fixé à l'extrémité libre de la corde et permet la fixation des poids du jeu d'apprentissage et de validation.

Cadran gradué. Secteur angulaire 0°-90° de rayons intérieur R_i et extérieur R_e, gradué au degré avec marquage accentué tous les 5° et 10°. Il permet une lecture directe de l'angle d'équilibre ϕ.

2.3 Principe de Fonctionnement

Le fonctionnement repose sur l'équilibre statique entre deux couples antagonistes autour du pivot O : le couple moteur et le couple résistant.
La masse m_{input} accrochée à la corde génère une tension sur la corde. Cette tension s'applique tangentiellement au tambour, produisant un couple moteur (équation 1). Le contrepoids et la masse du levier génèrent un couple résistant dépendant de l'angle ϕ (voir équation 1). La dépendance sinusoïdale introduit une non-linéarité dans le système mais acceptable pour des angles $\phi < 60°$. Le système atteint sa position d'équilibre ϕ_{eq} lorsque les deux couples se compensent exactement (equation 1)

$$\underbrace{m_{input} \cdot g \cdot R}_{\text{couple moteur}} = \underbrace{[m_{cw} \cdot r_{cw} + m_{lever} \cdot L/2] \cdot g \cdot sin(\phi_{eq})}_{\text{couple résistant}} \tag{1}$$

Cette équation admet une solution unique dans l'intervalle [0°, 90°] pour toute charge inférieure à la capacité maximale du système. La lecture de ϕ_{eq} sur le cadran constitue la mesure de sortie, directement reliée à la masse d'entrée.

3 Utilisation dans le cadre du "Neurone mécanique"

3.1 Procédures

Voici une procédure heuristique inspirée de la descente de gradient stochastique pour régression avec un jeu d'apprentissage de 5 échantillons. L'objectif est de trouver la position du contrepoids r_{cw} pour que l'angle mesuré ϕ corresponde aux angles cibles (les "vraies" valeurs). Le choix de 5 échantillons permet d'obtenir un temps d'apprentissage raisonnable. Ce nombre peut être augmenté si l'exercice porte davantage sur l'évaluation précise de l'erreur ou sur la capacité de généralisation du modèle.
0. Ce que l'élève a sous la main
   — Des poids (m_i) avec chacun une étiquette d'angle cible (t_i) (en degrés).
   — Le dispositif avec son cadran.
   — La grille à imprimer (figure 2) (une par époque).
   — Un crayon, une calculatrice
1. Choix des poids d'entraînement
L'élève choisit au hasard 5 poids dans le jeu de poids étiquetés pour l'entrainement et 5 poids pour la validation. La distance r_{cw} est placée aléatoirement sur le levier.
2. Règle de direction (signe du pas)
   — Calculez l'erreur pour chaque cas : $e_i = \phi_i - t_i$.
   — Calculez l'erreur moyenne $\bar{e} = \frac{1}{5}\sum_{i=1}^{5} e_i$.
3. Règle de taille du pas (10 mm, 5 mm ou 2 mm)
   — Calculez l'erreur absolue moyenne (MAE) : $\text{MAE} = \frac{1}{5}\sum_{i=1}^{5} |e_i|$

— Choisissez la taille du pas $|\Delta r|$ selon la MAE :
  — $MAE > 20° \Rightarrow 10\,\text{mm}$
  — $10° < MAE \leq 20° \Rightarrow 5\,\text{mm}$
  — $MAE \leq 10° \Rightarrow 2\,\text{mm}$

4. Algorithme (une itération = 1 "époque" sur 5 exemples)
   1. Position initiale : placez r_{cw} aléatoirement sur la glissière du levier.
   2. Mesures (pour $i = 1 \ldots 5$) :
      — placez m_i, laissez stabiliser, lisez l'angle ϕ_i
      — notez $e_i = \phi_i - t_i$.
   3. Agrégation : calculez $\bar{e}$ et la MAE.
   4. Décision : appliquez les règles de direction et de taille du pas
   5. Mise à jour : déplacez le contrepoids selon

$$r_{\text{cw}} \leftarrow r_{\text{cw}} + \text{sign}(\bar{e}) \times |\Delta r|$$

   6. Répétez les étapes 2 à 5 pendant 3 à 6 époques, jusqu'à obtenir $MAE \leq 2°$ de façon stable.

```
ÉPOQUE n° _____
Position r_cw (mm, repère sur la réglette) : _____

i   m_i (g)    t_i (°)     φ_i (°)     e_i = φ_i - t_i (°)
1   _______    _______     _______     _________________
2   _______    _______     _______     _________________
3   _______    _______     _______     _________________
4   _______    _______     _______     _________________
5   _______    _______     _______     _________________

Erreur moyenne signée  ȳ_e   = (Σ e_i) /5 = _______ °
Erreur absolue moyenne MAE = (Σ|e_i|)/5 = _______ °

Décision :
- Signe du pas (selon ȳ_e)      :  POSITIF / NÉGATIF
- Taille du pas (selon MAE)     : 10 mm / 5 mm / 2 mm

Nouvelle position r_cw = ____________________  (mm)
Commentaire (oscillation ? Frottement ?
saturation vers 90° ?) : ____________________
```

FIGURE 2 – Grille

À la fin de la séance, chaque groupe doit répondre à un QCM (Questionnaire à choix multiple) (tableau 1) pour évaluer les critères cognitifs sous forme d'objectifs pédagogiques mesurables.

4 Résultat

Le principe a été testé sur une population d'élèves du cycle 4 en classe de 4e et 3e du collège. Les élèves sont séparés en 6 groupes de 4 à 5 élèves. La séance dure au total 1 heure et 30 minutes. Trente minutes pour l'explication de l'activité et des notions puis une heure pour réaliser les manipulations.
La 1re activité de 30 min présente les bases de l'IA en définissant le domaine et en distinguant l'IA symbolique et l'apprentissage automatique. Il explique ensuite simplement le principe de l'apprentissage, fondé sur l'utilisation de données, d'un modèle et d'un algorithme d'optimisation, introduit la notion de fonction de coût, puis décrit de manière intuitive la descente de gradient stochastique (SGD). Enfin, l'utilisation du dispositif est présentée aux élèves.

TABLE 1 – QCM sur le neurone mécanique (énoncés condensés, sans réponses)

Notion	Énoncé
Entrée	Dans la balance à levier, quelle grandeur joue le rôle de l'entrée du neurone ?
Entrée	Dans un neurone artificiel, quel est le rôle de l'entrée x dans le calcul avec w ?
Poids	Quel composant correspond au poids synaptique w ?
Poids	Effet d'une augmentation de r_{cw} (à entrée constante) sur le couple et φ ?
Sortie	Quelle est la sortie mesurée du neurone mécanique (grandeur lue) ?
Erreur	Donner l'expression de l'erreur signée e_i en fonction de φ_i et t_i.
Erreur	Pour $MAE = 18°$, quel déplacement appliquer au contrepoids ?
Desc. grad.	Si l'erreur moyenne signée $\bar{e} < 0$ (sous-prédiction), comment ajuster r_{cw} ?
Desc. grad.	Pourquoi privilégier de petits pas successifs plutôt qu'un grand déplacement ?
Convergence	Quel critère indique la convergence de l'apprentissage ?
Convergence	Quel phénomène survient si le pas est trop grand ?

Dans notre échantillon, tous les groupes ont atteint le critère de convergence défini en 6 époques ou moins.
Les résultats du QCM sont analysés et présentés dans le tableau 2.

TABLE 2 – Statistiques descriptives par notion

Notion	Moyenne (0-1)	Écart-type	Σ des rangs (Friedman)
Entrée	0.833	0.258	28
Poids	0.417	0.376	15.5
Sortie	1.000	0.000	31
Erreur	0.500	0.316	16.5
Descente	0.417	0.376	15.5
Convergence	0.667	0.408	20

Les résultats préliminaires issus de l'étude pilote, associés à un test de Friedman, permettent les interprétations pédagogiques suivantes :

Certaines notions sont-elles trop faciles ? Oui : la notion "sortie" est clairement trop facile. Tous les indicateurs convergent : Moyenne = 1, écart-type = 0, somme des rangs le plus élevé de toutes les notions.

Certaines notions sont-elles trop difficiles ? Oui : deux notions ressortent comme difficiles (poids synaptique et descente de gradient). Elles ont les plus faibles moyennes, des écart-types élevés et des rangs faibles dans le test de Friedman.

Certaines notions sont-elles mal comprises ? Probablement oui. Les résultats variables entre groupes indiquent

une compréhension inégale.

Le QCM est-il équilibré ? Non. Une notion est trop facile, deux notions sont nettement plus difficiles, Les écarts-types sont très différents d'une notion à l'autre et le test de Friedman montre une tendance à des différences réelles $p \approx 0.05$.

Les différences observées sont-elles réelles ou dues au hasard ? Selon le test de Friedman (tableau 3), il existe une tendance réelle à des différences entre les notions. Ce n'est probablement pas dû au hasard.

TABLE 3 – Test global — Friedman (mesures appariées, k = 6 notions, n = 6 groupes)

Statistique Friedman : $\chi^2 = 11.036$
degrés de liberté : $df = 5$
p-value (exacte) : $p = 0.0503$

Cependant, les comparaisons post-hoc de Nemenyi ne révèlent aucune différence significative entre paires de notions (CD = 3.08). Cela suggère un QCM globalement cohérent, mais présentant des tendances de difficulté qui mériteraient d'être équilibrées dans une version future.

5 Discussion

Le dispositif mécanique proposé repose sur une analogie directe avec le neurone artificiel, dont le tableau 4 synthétise les correspondances terme à terme. Si cette analogie permet d'ancrer intuitivement les concepts fondamentaux de l'IA, elle présente néanmoins des limites inhérentes à sa nature physique.

TABLE 4 – Analyse et limite de l'analogie

Dispositif mécanique	Concept IA	Limite de l'analogie
Masse en entrée	Valeur d'entrée x	Une seule entrée possible
Position du contrepoids r_{cw}	Poids synaptique w	Pas de biais b modélisé
Angle ϕ	Sortie du neurone	Fonction $sin(\phi)$, pas sigmoïde
Déplacement du contrepoids	Mise à jour du gradient	Pas discret, non dérivable

La procédure d'ajustement manuel du contrepoids peut être interprétée comme une instance de descente de gradient stochastique (SGD). C'est une approximation discrète : le signe du déplacement correspond au signe du gradient, tandis que la taille du pas (10, 5 ou 2 mm) joue le rôle d'un taux d'apprentissage. Cette correspondance permet à l'apprenant de manipuler physiquement les deux décisions fondamentales de tout algorithme d'optimisation : la direction de la mise à jour et l'amplitude du pas.

Les résultats sont issus d'une étude pilote et devront être consolidés par une approche méthodologique plus rigoureuse. L'analyse entre les différentes notions abordées dans le QCM montrent que certaines notions sont mieux assimilées que d'autres. Les notions "entrée" et "sortie" sont rapidement comprises car elles ont un analogue physique direct et tangible, tandis que "descente de gradient" et "convergence" restent plus abstraites. Ces écarts montrent où accentuer les efforts pédagogiques. Ils permettent des améliorations dans la procédure et dans le QCM d'évaluation.

Enfin, si des travaux récents ont montré l'intérêt de faire « jouer le rôle de l'IA » aux apprenants à travers des manipulations symboliques (cartes ou jeu de rôle) [3], notre approche s'en distingue en proposant un analogue physique direct du neurone artificiel. La dynamique mécanique instancie exactement l'équation de pondération synaptique. Elle offre ainsi une approche intuitive par la manipulation (la position du contrepoids modifie l'équilibre) et formelle par la modélisation analytique (équation mathématique) absente des dispositifs purement symboliques. Et contrairement aux approches symboliques où l'apprenant simule le calcul, la contrainte physique du dispositif constitue une rétroaction sensorielle directe. Le levier ne ment pas : un mauvais réglage produit immédiatement un angle erroné.

Perspectives. À court terme, nous étendrons le dispositif vers un neurone à deux entrées en ajoutant un second tambour de rayon distinct et une seconde masse, de sorte que la somme des couples matérialise la combinaison linéaire. En parallèle, nous conduirons une évaluation de l'impact pédagogique plus structurée par pré-test / post-test sur un échantillon élargi, afin de mesurer les gains d'apprentissage sur les concepts cibles et d'estimer la robustesse des résultats (fiabilité du questionnaire, validation statistique).

6 Conclusion

Ce neurone mécanique montre qu'un dispositif simple et manipulable permet de rendre accessibles, sans recours au numérique, plusieurs principes fondamentaux de l'IA, notamment la relation entre entrée, poids synaptique et sortie ainsi que la logique d'un algorithme d'apprentissage inspiré de la descente de gradient. Les résultats de cette étude pilote suggèrent que les notions les plus ancrées dans le dispositif physique (comme l'entrée et la sortie) sont mieux comprises que les concepts plus abstraits (poids synaptique, descente de gradient, convergence), ce qui ouvre des pistes pour ajuster la scénarisation pédagogique et les supports. Ce dispositif ouvre aussi la voie à plusieurs extensions, notamment vers des neurones à entrées multiples ou vers l'ajout d'une came dont le profil permettra de simuler une activation sigmoïde et de ce fait, une régression logistique. Enfin, la création d'un jeu de données volontairement biaisés illustrera de manière expérimentale l'influence des biais présents dans les données sur la convergence et le résultat de l'apprentissage. Des évaluations plus méthodiques et sur une population plus large restent à mener pour confirmer l'impact didactique et pédagogique, et comparer son efficacité à d'autres approches d'"IA débranchée".

Références

[1] Cécile Barbachoux. Comprendre en agissant : effets de l'approche STEM sur l'apprentissage des grandeurs et mesures en mathématiques au cycle 3. working paper or preprint, April 2025.

[2] Maria Cristina Carrisi, Mirko Marras, and Sara Vergallo. A structured unplugged approach for foundational ai literacy in primary education, 2025.

[3] Hasti Darabipourshiraz, Lily Murakami Ng, Grace Wang, Sophie Rollins, and Duri Long. Ai unplugged : Exploring pathways from physical simulation to conceptualization of ai reasoning processes. *ACM Trans. Comput. Educ.*, January 2026. Just Accepted.

[4] Fengchun Miao, Wayne Holmes, Ronghuai Huang, and Hui Zhang. *AI and Education : Guidance for Policy-Makers*. UNESCO, Paris, 2021.

[5] Micha Mueller, Annabel Lindner, Patrick Löffler, and Marc Berges. Das wollperzeptron - eine unplugged aktivität zum maschinellen lernen. *Informatische Bildung in Schulen*, 3, 09 2025.

[6] W.S. Orr. *Orr's Circle of the Sciences : Mechanical philosophy (1856)*. Orr's Circle of the Sciences : A Series of Treatises on the Principles of Science, with Their Application to Practical Pursuits. W.S. Orr and Company, 1856.

[7] Daniel Raviv and Daniel Barb. A visual and engaging approach to teaching and learning the normal distribution. *Journal of Higher Education Theory and Practice*, 20(14), Dec. 2020.

[8] Axel Schaffland and Julius Schöning. Mechanical neural network : Making ai comprehensible for everyone. In *IEEE Global Engineering Education Conference (EDUCON)*, pages 1–6, 08 2023.

[9] Mira Sneirson, Josephine Chai, and Iris Howley. A learning approach for increasing ai literacy via xai in informal settings. In *Artificial Intelligence in Education*, pages 336–343. Springer Nature Switzerland, 2024.

[10] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. Envisioning ai for k-12 : What should every child know about ai ? *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 :9795–9799, 07 2019.

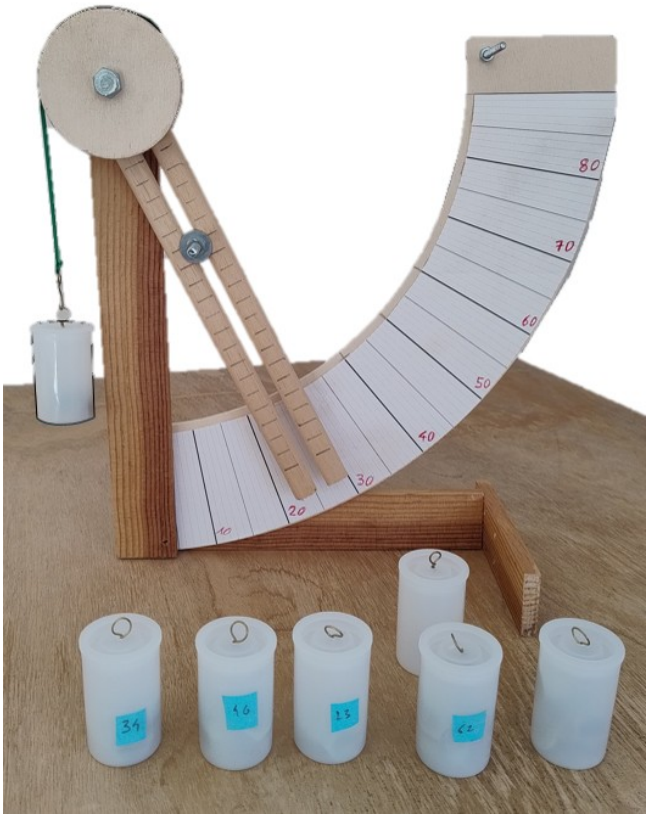


FIGURE 3 – Photo du dispositif avec les données d'apprentissage

SESSION 4 : Explicabilité et apprentissage de structure présidée parIsmaïl Baaj

Apprentissage de dynamiques latentes. Stabilisation par structure hamiltonienne

Nicolas Gutowski, Jordan Levy, Wajdi Dhifli, Bounia Louenas

Apprentissage de dynamiques latentes : stabilisation par structure hamiltonienne

F. Lionti[1], N. Gutowski[2], S. Aubin[3], P. Martinet[1]

[1] Université de Côte d'Azur, INRIA
[2] Université d'Angers, LERIA
[3] Direction Générale de l'Armement - Techniques Terrestres

Résumé

L'apprentissage de modèles physiques à partir de données est un problème de longue date en identification de systèmes et en apprentissage automatique. Garantir une généralisation fiable sur le domaine d'intérêt est crucial pour le déploiement sécurisé des modèles appris dans les applications de robotique et d'ingénierie. Ce défi est accentué dans les contextes impliquant des horizons de prédiction à long terme et des variables d'état partiellement observées, où les représentations latentes non contraintes ont tendance à accumuler des erreurs au fil du temps. Dans ce contexte, nous proposons une architecture neuronale hamiltonienne pour l'apprentissage de systèmes dynamiques partiellement observés. En exploitant un biais inductif de conservation de l'énergie, notre approche combine un encodeur de contexte avec une formulation hamiltonienne, permettant l'inférence de l'état latent à partir d'observations partielles tout en préservant la structure physique sous-jacente. Le modèle est entraîné sur des fenêtres d'observation courtes, bruitées et partielles, puis évalué sur des prédictions à long terme afin de mesurer la stabilité et la robustesse de sa généralisation. Les expériences menées sur des variantes non amorties et conservatives des systèmes de Duffing, du pendule, de Hénon-Heiles et de Kepler démontrent une amélioration significative de la stabilité à long terme et de la précision de prédiction par rapport aux modèles de référence non structurés.

1 Introduction

L'apprentissage de modèles dynamiques à partir de données est au cœur de la modélisation, du contrôle pour la robotique et l'ingénieurie [10]. Pour les applications critiques en matière de sécurité, une précision sur de courtes échelles de temps ne suffit pas, les modèles doivent fournir des prédictions fiables à long terme, là où de petites erreurs de modélisation peuvent s'accumuler, entraînant une dérive catastrophique. Cette sensibilité a déplacé l'attention des régresseurs traditionnels en temps discret vers des formulations en temps continu [9] pour leurs propriétés de robustesse accrues. En assimilant les mises à jour résiduelles à des schémas de discrétisation d'équations différentielles, les Neural ODEs se sont imposées comme un cadre robuste pour l'apprentissage de champs de vecteurs paramétrés [1]. Bien que ces modèles offrent une meilleure cohérence temporelle, ils restent fondamentalement de nature "boîte noire", car ils n'imposent pas explicitement d'invariants physiques ou de quantités conservées.

Pour combler cette lacune, des architectures informées par la physique telles que les Réseaux de Neurones Hamiltoniens (HNNs) [4] et les Réseaux de Neurones Lagrangiens (LNNs) [11] ont été proposées. En intégrant des biais inductifs basés sur l'énergie directement dans l'architecture du réseau, ces modèles préservent la structure géométrique sous-jacente du flot, garantissant la stabilité à long terme et la conservation de l'énergie. Cependant, leur utilité pratique est souvent limitée par une exigence stricte : ils supposent généralement l'accès à l'état canonique complet du système (par exemple, les positions et les moments). Dans les contextes d'ingénierie réels, les capteurs ne fournissent que des observations partielles, rendant ces approches structurées inapplicables sans une méthode de reconstruction des dimensions manquantes de l'espace des phases.

Dans le régime d'observation partielle, l'état du système doit être inféré par des représentations latentes. Pourtant, lorsque ces espaces latents ne sont pas contraints, le modèle est libre d'apprendre des transitions mathématiquement pratiques mais physiquement impossibles. Ce manque d'ancrage structurel provoque une dérive de la dynamique latente, menant à des trajectoires qui violent des lois fondamentales, comme la conservation de l'énergie [8]. Dans les tâches à long horizon, des petites incohérences dans le processus d'inférence latente s'amplifient avec le temps, entraînant des comportements instables, dissipatifs ou physiquement peu plausibles qui compromettent la fiabilité du modèle appris.

Pour remédier à ces limitations, nous proposons un cadre qui impose une structure hamiltonienne dans un espace latent appris. Plutôt que de traiter l'état latent comme un vecteur de caractéristiques arbitraire, nous le contraignons à évoluer comme un ensemble de coordonnées structurées régies par un flot hamiltonien, fournissant une représentation de la dynamique sous-jacente respectueuse de la géométrie symplectique. Nos principales contributions sont les suivantes :

— Nous proposons un cadre de modélisation dynamique latente qui impose une structure hamiltonienne dans un espace latent appris, utilisant des contraintes géométriques comme biais inductif pour

favoriser des flots conservatifs et une stabilité à long terme sous observabilité partielle.
— Nous introduisons un encodeur basé sur le contexte qui infère les conditions initiales latentes à partir de fenêtres d'observation courtes et bruitées, permettant une reconstruction efficace de l'espace des phases sans accès aux états canoniques complets.
— Nous fournissons une analyse théorique et empirique (App. A) caractérisant comment la longueur du contexte, le bruit et la discrétisation induisent conjointement une incertitude d'initialisation irréductible, et comment les projections symplectiques préservant la structure atténuent l'amplification de cette incertitude au fil du temps.
— À travers des expériences approfondies sur des systèmes dynamiques non linéaires, nous démontrons que la structure hamiltonienne réduit systématiquement l'accumulation d'erreurs à long terme et la variabilité d'un entraînement à l'autre par rapport aux modèles de référence Neural ODE latents non contraints.

1.1 Travaux antérieurs

La prédiction précise à long terme demeure un défi persistant pour les modèles de systèmes dynamiques appris à partir de données. Même lorsque les erreurs à un pas de temps sont faibles, la répétition des prédictions peut amplifier les incohérences numériques et de modélisation, entraînant des trajectoires instables et des violations d'invariants physiques. Les travaux antérieurs ont abordé ce problème selon trois axes complémentaires (i) les formulations en temps continu qui améliorent la cohérence temporelle, (ii) les biais inductifs préservant la structure pour encoder la physique connue, et (iii) les mécanismes d'inférence d'états latents pour les systèmes partiellement observés. Nous passons en revue ces axes et mettons en évidence une lacune à leur intersection : garantir la stabilité des prédictions à long terme sous observabilité partielle, tout en préservant la structure hamiltonienne.

Neural ODEs et prédictions longues. Les méthodes basées sur les *Neural Ordinary Differential Equations* (Neural ODEs) [1] apprennent des dynamiques en temps continu en paramétrant un champ de vecteurs, lequel est intégré numériquement pour correspondre aux trajectoires observées. Au-delà de l'objectif d'entraînement standard à un pas, les méthodes d'entraînement multi-pas montrent une robustesse accrue au bruit et aux erreurs cumulatives lorsque les modèles sont optimisés sur des trajectoires complètes [9]. L'entraînement multi-pas favorise l'apprentissage de dynamiques stables par rapport à la prédiction étape par étape. Cependant, les Neural ODEs restent largement non contraintes : le champ de vecteurs appris n'est pas tenu de strictement satisfaire aux lois de conservation ou aux propriétés géométriques du système sous-jacent. Par conséquent, les projections à long terme peuvent encore dériver, en particulier sous observabilité partielle ou en présence de bruit d'observation, là où l'état inféré peut évoluer de manière physiquement incohérente.

Dynamiques préservant la structure. Une approche complémentaire consiste à incorporer des biais inductifs qui préservent la structure physique connue. Les *Hamiltonian Neural Networks* (HNN) [4] et les *Lagrangian Neural Networks* (LNN) [11, 2] paramètrent une fonction d'énergie (ou d'action) dont ils dérivent la dynamique via les équations de Hamilton ou d'Euler-Lagrange. Ce formalisme favorise le respect des invariants physiques et garantit une meilleure stabilité à long terme. D'autres approches s'appuient sur des fonctions de Lyapunov [12] pour contraindre la stabilité, visant ainsi à contrôler la dérive de l'énergie et à préserver la cohérence des prédictions multi-pas sur le long terme. En pratique, cependant, beaucoup de ces méthodes supposent l'accès aux coordonnées canoniques (par ex., (q, p)) ou à des mesures d'état suffisamment riches, ce qui limite leur applicabilité directe aux systèmes réels où les capteurs ne fournissent que des observations partielles.

Apprentissage sous observabilité partielle. Lorsque l'état complet n'est pas observé, les modèles introduisent typiquement un état latent inféré à partir de séquences d'observations à l'aide d'encodeurs ou de modèles récurrents [5, 7]. Dans ce contexte, les Neural ODEs latents [13] étendent l'apprentissage en temps continu à ce cadre en couplant un réseau d'inférence avec un modèle de transition latent, permettant l'apprentissage de la dynamique à partir de mesures irrégulières et incomplètes. Cependant, les transitions latentes sont non contraintes, ainsi, rien n'empêche les dynamiques inférées de violer les lois de conservation, et de petites erreurs d'inférence peuvent s'accumuler en une dérive substantielle à long horizon. Les approches préservant la structure sous observabilité partielle reposent souvent sur l'apprentissage de plongements latents à haute dimension, transformant les observations brutes vers des formes canoniques sans supposer l'accès préalable aux coordonnées physiques du système [14], mais elles ne traitent pas directement la stabilité à long terme des projections latentes lorsque l'inférence est effectuée à partir de fenêtres courtes et bruitées.
Notre travail se situe à cette intersection : nous imposons une structure hamiltonienne directement au sein des dynamiques latentes inférées. Cette approche est couplée à une reconstruction de l'espace des phases exploitant des plongements à retard (*delay embeddings*) appris via un encodeur de type *Gated Recurrent Unit* (GRU).

2 Méthode

Afin d'aborder le problème de l'apprentissage de dynamiques stables à long terme à partir d'observations partielles et bruitées, notre approche repose sur trois piliers : (i) la reconstruction de l'espace des phases via un encodeur de contexte, (ii) une formulation hamiltonienne des dynamiques latentes, et (iii) une intégration numérique préservant la structure et robuste au bruit d'observation.

Reconstruction de l'espace des phases. Nous considérons un système déterministe où l'état complet $x(t) \in \mathbb{R}^{d_x}$ est régi par $\dot{x}(t) = F(x(t))$. Nous n'avons accès qu'à une observation partielle $y(t) \in \mathbb{R}^{d_y}$, souvent corrompue par du

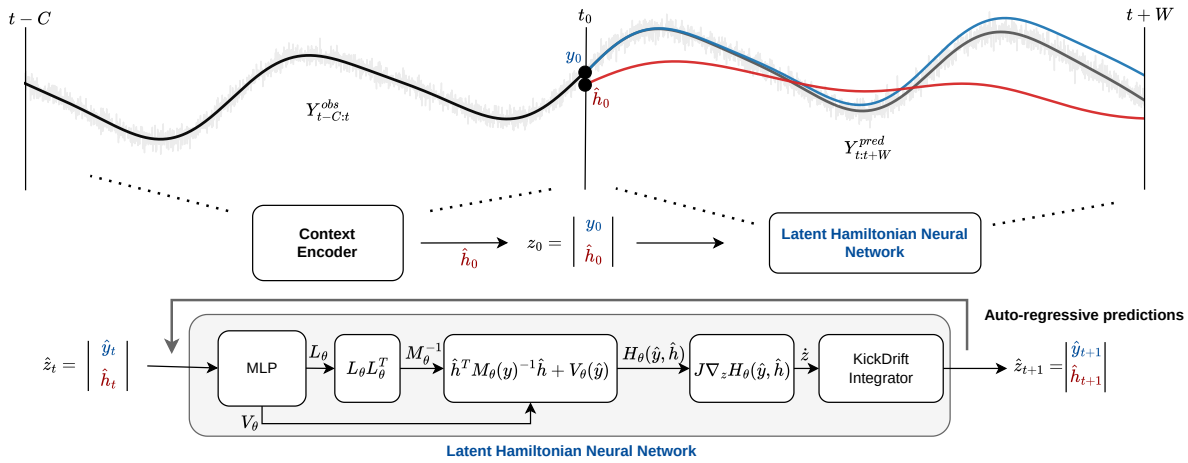


FIGURE 1 – Aperçu du cadre de prévision Hamiltonien latent proposé. Une fenêtre d'observations passées de longueur C est encodée en un état latent initial h_0 par un encodeur de contexte. Le Réseau de Neurones Hamiltonien latent modélise la dynamique latente en temps continu, laquelle est intégrée numériquement à l'aide d'un intégrateur symplectique de type « kick–drift » de premier ordre pour produire les trajectoires latentes futures sur un horizon W. Les prédictions observables $\hat{y}_{t:t+W}$ sont obtenues via un processus de décodage auto-régressif dans l'espace des observations, tandis que les variables latentes évoluent sous une structure hamiltonienne explicite qui capture les propriétés dynamiques sous-jacentes du système.

bruit : $y^{\text{obs}}(t) = y(t) + \varepsilon(t)$. Nous utilisons un encodeur de contexte E_ϕ, implémenté sous forme de réseau de neurones récurrent de type GRU, pour projeter un historique de C observations $Y^{\text{obs}}_{t-C:t}$ vers une prédiction d'état latent initial $\hat{h}_0 \in \mathbb{R}^{d_h}$. Nous formons ensuite l'état augmenté $z_0 = [y_0, \hat{h}_0]^\top$ en concaténant la prédiction latente initiale avec une observation initiale. Dans nos expériences, nous imposons $d_h = d_y$ afn de faciliter le couplage symplectique.

Dynamique Hamiltonienne latente riemannienne. Pour garantir la stabilité à long terme, nous contraignons l'évolution de z_t à un flot hamiltonien. Nous paramétrons une fonction hamiltonienne latente $H_\theta(y, h)$ à l'aide d'un réseau de neurones structuré :

$$H_\theta(y, h) = \frac{1}{2} h^\top M_\theta(y)^{-1} h + V_\theta(y), \tag{1}$$

où V_θ représente l'énergie potentielle et $M_\theta(y)$ est une matrice de masse apprise dépendant de la position. Cette formulation riemannienne permet au modèle d'apprendre des propriétés inertielles dépendantes de la position et généralise les énergies cinétiques euclidiennes classiques. Pour garantir que $M_\theta(y)$ soit toujours définie positive symétrique (SPD), nous paramétrons son inverse via un facteur de Cholesky $L_\theta(y)$:

$$M_\theta(y)^{-1} = L_\theta(y) L_\theta(y)^\top, \tag{2}$$

où L_θ est une matrice triangulaire inférieure avec une diagonale strictement positive imposée par une transformation softplus [3]. Cette formulation riemannienne permet au modèle d'apprendre des structures cinétiques complexes et non linéaires dans l'espace latent. Le champ de vecteurs est ensuite dérivé du crochet de Poisson :

$$\dot{z} = J\nabla_z H_\theta(z), \quad J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}. \tag{3}$$

Intégration préservant la structure. Nous utilisons un intégrateur symplectique de premier ordre (kick–drift / Euler symplectique) [6] pour projeter la dynamique hamiltonienne apprise. Pour un pas de temps Δt, la mise à jour est :

$$h_{k+1} = h_k - \Delta t \nabla_y H_\theta(y_k, h_k) \tag{4}$$

$$y_{k+1} = y_k + \Delta t \nabla_h H_\theta(y_k, h_{k+1}) \tag{5}$$

Ce choix préserve la forme symplectique au niveau discret, ce qui empêche la dissipation artificielle ou l'injection d'énergie pouvant survenir avec des solveurs non symplectiques, et réduit considérablement la dérive à long terme. Il est également efficace sur le plan computationnel : chaque étape ne nécessite que deux évaluations de gradient et s'avère moins coûteuse que les schémas symplectiques de second ordre courants (ex. Störmer–Verlet/leapfrog), ce qui le rend bien adapté à l'entraînement et aux projections longues.

Objectif d'entraînement par fenêtre. Le modèle (θ, ϕ) est entraîné de bout en bout pour minimiser l'erreur quadratique moyenne sur une fenêtre de prédiction W :

$$\mathcal{L}(\theta, \phi) = \mathbb{E}\left[\frac{1}{W}\sum_{k=0}^{W-1} \|\hat{y}_{t+k} - y^{\text{obs}}_{t+k}\|_2^2\right], \tag{6}$$

où $\hat{y}$ est la prédiction de l'observation prédite. L'entraînement sur des fenêtres bruitées y^{obs} force l'encodeur à extraire une condition initiale latente robuste, tandis que la contrainte hamiltonienne garantit que la prédiction suivante reste physiquement plausible.

Robustesse avec état latent initial bruité. Nous fournissons une analyse théorique (Annexe A) expliquant comment l'observabilité partielle, la longueur du contexte et la discrétisation numérique déterminent conjointement la robustesse

à long terme. Notre idée clé est que l'erreur de prédiction à long horizon est dominée par l'amplification de l'incertitude résiduelle dans l'état latent initial inféré, plutôt que par l'erreur locale de prédiction à un pas.

Sous observabilité partielle, le moment latent initial $\hat{h}_0$ doit être inféré à partir d'une fenêtre de contexte finie et bruitée $Y^{\text{obs}}_{t-C:t}$. L'encodeur GRU agit comme un filtre temporel : augmenter la longueur du contexte C permet d'agréger des mesures redondantes et d'améliorer le rapport signal/bruit de $\hat{h}_0$. Cependant, comme les prédictions sont ancrées à la coordonnée initiale observée y^{obs}_0, le bruit de mesure induit une contribution irréductible à la covariance initiale $\Sigma_0 = \text{Cov}([y^{\text{obs}}_0, \hat{h}_0]^\top)$. Par conséquent, augmenter C ne réduit la variance de $\hat{h}_0$ que jusqu'à un plancher d'incertitude induit par le bruit qui ne peut être éliminé par la seule inférence latente.

Une linéarisation au premier ordre de la dynamique latente montre que cette incertitude résiduelle est propagée et amplifiée au cours du temps par l'application répétée de la fonction de transition. En conséquence, même de modestes différences dans la qualité de l'initialisation peuvent entraîner des écarts de performance substantiels à long terme. Crucialement, le choix de la discrétisation numérique affecte directement cette amplification. Alors que les schémas non symplectiques peuvent introduire des artefacts dissipatifs ou expansifs fallacieux qui amplifient artificiellement l'incertitude, les prédictions symplectiques préservant la structure maintiennent la rigidité géométrique du flot hamiltonien et limitent l'amplification numérique.

Ensemble, cette analyse explique les compromis empiriques observés entre la longueur du contexte C, l'horizon de prédiction W et la stabilité à long terme, et motive la conception conjointe de la structure hamiltonienne latente et de la discrétisation symplectique.

3 Cadre expérimental

Nous focalisons l'évaluation de notre méthode sur ses capacités de prédiction à long horizon, dans un contexte d'observation partielle corrompu par du bruit, pour tester la robustesse aux erreurs cumulatives. Dans ce contexte, les modèles doivent inférer des moments non observés à partir de courtes fenêtres de configurations bruitées et produire des prédictions stables sur de longs horizons. Nous rapportons les performances en utilisant la RMSE à long horizon sur les données de test sur 3 graines (seeds) et analysons l'impact de la longueur du contexte C et de l'horizon de prédiction W.

Données, observabilité partielle et bruit. Les expériences sont menées sur quatre systèmes dynamiques non linéaires : Duffing, le Pendule, Hénon–Heiles (noté *Henon*), et Kepler. Duffing et le Pendule sont des systèmes 2D avec des états $x = (q, \dot{q})$ et $x = (\theta, \omega)$, respectivement, tandis que Hénon–Heiles et Kepler sont des systèmes hamiltoniens 4D avec des coordonnées canoniques $x = (q_1, q_2, p_1, p_2)$. Toutes les trajectoires sont générées par simulation numérique avec un pas de temps fixe $\Delta t = 0.02$ sur $T = 500$ pas de temps. Pour chaque système, nous générons 256 trajectoires d'entraînement, 64 de validation et 64 de test, chacune à partir de conditions initiales aléatoires. Nous considérons à la fois des observations propres et des observations bruitées, dans le cadre bruité, un bruit gaussien i.i.d. d'écart-type $\sigma = 0.02$ est ajouté aux variables observées pour tous les systèmes. Le problème d'apprentissage est partiellement observé : le modèle ne reçoit que les variables de configuration (par ex., q pour Duffing/Pendule, (q_1, q_2) pour Hénon–Heiles et Kepler), tandis que les variables de moment ne sont pas observées et doivent être inférées. Les équations dynamiques complètes et les distributions des conditions initiales utilisées pour la génération des données sont fournies en App. C.

Pourquoi observer uniquement q ? Nous choisissons un modèle d'observation portant uniquement sur la position (configuration seule) car il constitue le régime d'inférence le plus difficile et fournit donc un test de résistance significatif pour la robustesse à long horizon. Récupérer les positions à partir des vitesses/moments est une opération de lissage (intégration), alors que récupérer les vitesses/moments à partir des positions nécessite une différenciation, laquelle est mal conditionnée et amplifie fortement le bruit de mesure. Ainsi, observer q force le modèle à effectuer une estimation de moment implicite et robuste au bruit à partir de fenêtres courtes, plutôt que de s'appuyer sur une application inverse plus facile. Ce cadre met en évidence si une méthode peut reconstruire de manière stable les composantes manquantes de l'espace des phases et maintenir des prédictions physiquement plausibles sous observabilité partielle.

Entraînement et évaluation à long horizon. L'entraînement est effectué sur des fenêtres glissantes. Pour chaque trajectoire, nous extrayons des échantillons composés d'un contexte passé $y_{t-C:t}$ de longueur C et d'une fenêtre de prédiction future $y_{t:t+W}$ de longueur W, en utilisant un pas (stride) de 10. Lors de l'évaluation, les prédictions du modèle sont obtenues de manière auto-régressive sur un long horizon de $L = 200$ pas, en se conditionnant sur le contexte observé et en mesurant la RMSE par rapport à l'état réel (ground-truth) non bruité.

Nous comparons trois méthodes : (i) **Latent NODE (E)**, un champ latent non contraint intégré avec Euler explicite, (ii) **Latent HNN (E)**, un modèle à paramétrage hamiltonien utilisant Euler explicite ; et (iii) **Latent HNN (S)**, utilisant un schéma symplectique (kick–drift). Bien que les réseaux puissent compenser partiellement les erreurs de discrétisation, les solveurs non symplectiques échouent à apprendre les lois de conservation car leur Inverse Modified Differential Equations (IMDE) n'est pas hamiltonienne [15]. Les états initiaux sont inférés via un encodeur GRU. Entraînés pendant 150 époques avec Adam (10^{-4}) et un écrêtage de gradient (gradient clipping) de 1.0, les meilleurs points de sauvegarde (checkpoints) minimisent la RMSE de validation à long horizon ($L = 200$).

Sur la complexité des solveurs et les gradients d'entraînement. Les solveurs d'ordre supérieur offrent des avantages limités car l'entraînement des Neural ODEs approxime une IMDE spécifique au solveur plutôt que l'ODE réelle [15]. Les réseaux à haute capacité absorbent efficacement les er-

reurs de discrétisation, mais cette compensation manque de cohérence physique pour les systèmes hamiltoniens, à moins que l'intégrateur ne soit symplectique. Par souci d'efficacité, nous utilisons une discrétisation à pas fixe et différencions à travers le graphe de calcul discret, en privilégiant la stabilité et la vitesse d'entraînement plutôt que l'ordre formel de discrétisation.

4 Résultats

Dans cette section, nous évaluons les performances de notre méthode par rapport aux modèles de référence non contraints. Notre analyse se concentre sur trois aspects clés : la stabilité des trajectoires à long terme, la conservation de l'énergie et la robustesse de la reconstruction de l'espace des phases sous observabilité partielle.

Résumé des performances à long horizon. Le Tableau 1 rapporte la RMSE à long horizon pour une longueur de projection (rollout) fixe $L = 200$ pour quatre systèmes sous observations propres et bruitées ($\sigma = 0,02$). Pour chaque modèle, le couple (C, W) est sélectionné en minimisant la RMSE moyenne de validation à long horizon sur plusieurs graines (seeds), et nous rapportons la RMSE de test correspondante (moyenne $\pm$ std). Dans l'ensemble, imposer un paramétrage hamiltonien améliore la stabilité à long terme par rapport à la méthode Latent NODE. L'utilisation d'un intégrateur symplectique renforce encore cet effet en empêchant la dérive induite par l'intégration. Le bénéfice est plus marqué sur les systèmes de Kepler et de Hénon–Heiles, où le modèle HNN (S) permet des réductions d'erreurs importantes et une plus faible variabilité entre les graines dans le cadre bruité, indiquant une meilleure robustesse aux erreurs cumulatives sous observabilité partielle. Sur le Pendule, le modèle HNN (E) est moins performant que Latent NODE, soulignant qu'un modèle hamiltonien, sans une discrétisation préservant cette structure, peut être préjudiciable, le passage au modèle HNN (S) rétablit des gains constants, confirmant l'importance de la cohérence entre le modèle et l'intégrateur.

La Figure 2 apporte une interprétation complémentaire en visualisant l'évolution de la RMSE (erreur quadratique moyenne) de test sur des prédictions à long terme (ou rollouts) pour chaque système. Le système du pendule met en évidence l'importance de la cohérence de l'intégration numérique et de la contrainte symplectique dans la dynamique latente. Bien que le paramétrage hamiltonien puisse en principe améliorer la stabilité, le modèle HNN (E) ne surpasse pas systématiquement le modèle de référence non structuré. En revanche, le modèle HNN (S) atteint une erreur plus faible et une variance réduite sur une large gamme de tailles de fenêtres, ce qui indique que la contrainte hamiltonienne seule est insuffisante à moins que le schéma d'intégration numérique ne préserve la structure géométrique sous-jacente. Ce comportement s'étend de manière cohérente à des dynamiques plus complexes, soulignant davantage le rôle de contraintes structurelles fortes dans la stabilisation des représentations latentes.

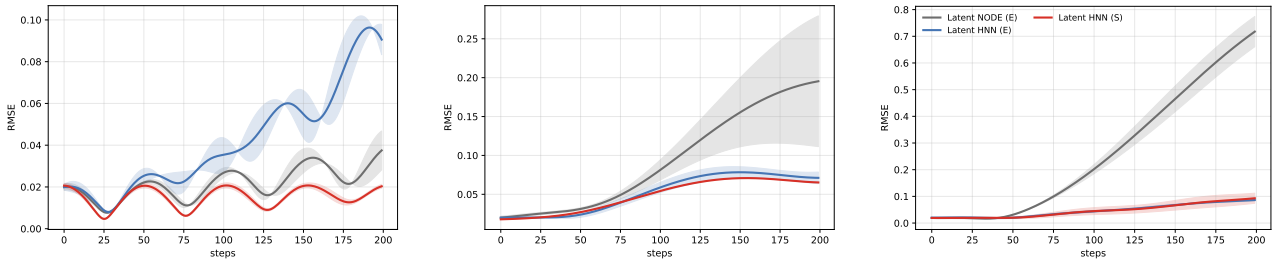
Sur le système de Hénon–Heiles, les modèles latents non structurés présentent des trajectoires incohérentes selon les graines aléatoires, ce qui entraîne une variance élevée et une croissance rapide de l'erreur à mesure que l'horizon de prédiction augmente. Les modèles basés sur le formalisme hamiltonien imposent une contrainte forte sur la dynamique latente qui régularise son évolution, aboutissant à des trajectoires plus cohérentes, une variance réduite et une meilleure précision à long horizon. La variante symplectique renforce encore cet effet en préservant la structure géométrique apprise lors de l'intégration numérique.

Une tendance similaire est observée sur le système de Kepler. Le modèle de référence Latent NODE montre à nouveau des prédictions instables et très variables selon les graines, avec une accumulation d'erreurs sévère à mesure que la longueur du contexte augmente. Les modèles hamiltoniens atténuent considérablement ces effets en imposant une dynamique latente structurée, et le couplage de cette structure avec une intégration symplectique produit les prédictions à long horizon les plus stables et les plus robustes.

Ensemble, ces résultats confirment que des contraintes géométriques fortes, combinées à une intégration numérique cohérente avec la structure, sont essentielles pour réduire la variabilité entre les exécutions et maintenir la précision sur des horizons étendus sous observabilité partielle.

	Duffing				**Henon**			
	Clean		Noisy		Clean		Noisy	
Model	RMSE ↓	Gain ↑	RMSE ↓	Gain ↑	RMSE ↓	Gain ↑	RMSE ↓	Gain ↑
Latent NODE	0.022 ± 0.000	–	0.035 ± 0.002	–	0.038 ± 0.005	–	0.112 ± 0.034	–
Latent HNN (E)	0.007 ± 0.000	$+214$	0.029 ± 0.003	$+21$	0.037 ± 0.006	$+3$	0.057 ± 0.004	$+96$
Latent HNN (S)	0.004 ± 0.000	$+450$	0.029 ± 0.003	$+21$	0.038 ± 0.003	$+0$	0.052 ± 0.000	$+115$
	Pendulum				**Kepler**			
	Clean		Noisy		Clean		Noisy	
Model	RMSE ↓	Gain ↑	RMSE ↓	Gain ↑	RMSE ↓	Gain ↑	RMSE ↓	Gain ↑
Latent NODE	0.009 ± 0.000	–	0.023 ± 0.001	–	0.308 ± 0.000	–	0.349 ± 0.032	–
Latent HNN (E)	0.043 ± 0.000	-79	0.048 ± 0.002	-52	0.033 ± 0.000	$+835$	0.052 ± 0.000	$+575$
Latent HNN (S)	0.003 ± 0.000	$+200$	0.016 ± 0.001	$+44$	0.029 ± 0.000	$+957$	0.052 ± 0.013	$+568$

TABLE 1 – RMSE de test (moyenne $\pm$ écart-type sur les graines aléatoires) pour des observations propres et bruitées ($\sigma = 0,02$) ; pour chaque modèle, le couple (C, W) est sélectionné sur l'ensemble de validation. Le gain est relatif au modèle Latent NODE (plus il est élevé, mieux c'est). Les résultats détaillés et les balayages complets de (C, W) sont rapportés en Annexe E.



(a) Pendulum (b) Hénon–Heiles (c) Kepler

FIGURE 2 – RMSE à long horizon ($L = 200$) sous observations bruitées en fonction des paramètres de fenêtre. Le paramètre complémentaire est sélectionné sur l'ensemble de validation : le meilleur C pour chaque W et le meilleur W pour chaque C. Voir l'Annexe E pour les détails expérimentaux complets.

Taille du contexte et horizon d'entraînement. La Figure 3 rapporte une analyse de sensibilité sur le système Hénon-Heiles sous des observations bruitées, mettant en lumière les rôles complémentaires de la longueur de contexte C et de l'horizon de déploiement W dans la robustesse à long horizon. À mesure que W augmente, la RMSE à long horizon et sa variance se stabilisent progressivement pour tous les modèles. Ce comportement est cohérent avec la décomposition biais–variance des projections à long terme [9] : faire la moyenne de la perte sur des trajectoires plus longues réduit la variance de l'objectif d'entraînement jusqu'à atteindre un régime asymptotique pour les systèmes dynamiques non expansifs. Dans notre configuration, cette condition est satisfaite car la dynamique hamiltonienne, combinée à l'intégration symplectique, préserve le volume de l'espace des phases et évite l'amplification artificielle de l'incertitude (voir App. A). Néanmoins, un plancher de variance subsiste même pour de grandes valeurs de W, reflétant l'incertitude irréductible induite par la coordonnée initiale observée et bruitée y_0^{obs}. Puisque les prédictions sont ancrées à cette observation, le bruit de mesure induit une contribution non nulle à la covariance initiale qui ne peut être éliminée en étendant l'horizon de projection (App. A). Au-delà de ce point, augmenter W ne réduit plus la variance de la prédiction.

Une fois que le signal de supervision est stabilisé par un W suffisamment grand, l'augmentation de la longueur du contexte C améliore la robustesse en permettant à l'encodeur de mieux exploiter ce signal d'entraînement. Des fenêtres de contexte plus longues renforcent le filtrage temporel effectué par l'encodeur, menant à une inférence plus cohérente du moment latent initial $\hat{h}_0$ à travers les différentes graines aléatoires. Empiriquement, cet effet se manifeste principalement par une réduction de la variance de la RMSE à long horizon plutôt que par une diminution supplémentaire de sa moyenne, indiquant une meilleure robustesse de l'inférence latente sous l'effet du bruit. Dans l'ensemble, ces résultats démontrent que la robustesse à long horizon provient de l'interaction entre des prédictions symplectiques stabilisées sur W et un filtrage latent efficace sur C, tous deux opérant sous un seuil d'incertitude irréductible fixé par l'observation initiale bruitée.

Géométrie de l'espace des phases et jauge d'énergie. L'hamiltonien appris définit une jauge d'énergie latente, c'est-à-dire qu'il est défini à une échelle et un décalage près. Dans ce contexte, la Figure 4 fournit des diagnostics géométriques complémentaires à travers les trajectoires de l'espace des phases latent. Nous soulignons que l'Hamiltonien appris

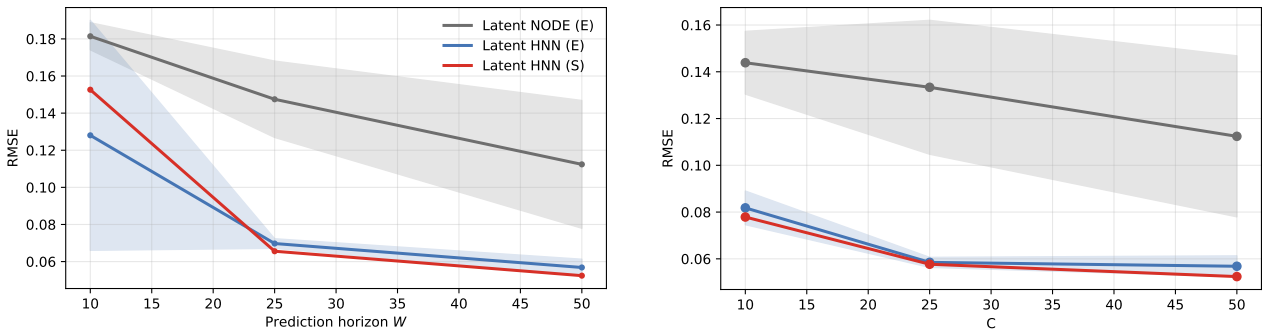


FIGURE 3 – Chaque graphique présente la RMSE de test obtenue en sélectionnant le paramètre de fenêtre complémentaire optimal (meilleur C pour chaque W à gauche, et meilleur W pour chaque C à droite) sur le système de Hénon-Heiles sous observations bruitées.

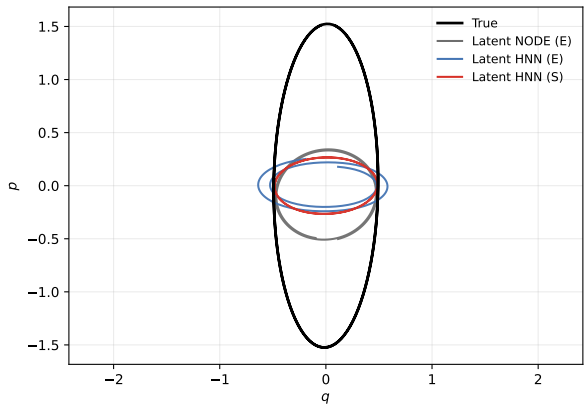


FIGURE 4 – Trajectoires dans l'espace des phases latent pour le système du pendule sous observations bruitées (meilleur (C, W)). Le modèle HNN (S) symplectique préserve une structure orbitale cohérente, contrairement aux variantes non symplectiques. L'hamiltonien appris définit une jauge d'énergie latente : l'échelle absolue des orbites n'a pas de signification physique.

H_θ ne doit pas être interprété comme la véritable énergie physique. Il est défini à un choix de jauge près et sert de coordonnée d'énergie latente plutôt que de quantité physique absolue. Par conséquent, la taille et l'échelle des orbites dans l'espace des phases ne sont pas significatives en soi d'un point de vue physique.

Malgré cette liberté de jauge, les trajectoires dans l'espace des phases révèlent clairement l'effet des contraintes symplectiques. Le modèle HNN (S) symplectique produit des orbites cohérentes et fermées qui préservent la géométrie qualitative de la dynamique, alors que la NODE non contrainte et le modèle Hamiltonien couplé à une discrétisation non symplectique présentent une contraction ou une expansion artificielle des trajectoires. Différents choix de jauge (par exemple, induits par la matrice de masse apprise) peuvent redimensionner les orbites, mais ils n'affectent pas la préservation de la structure symplectique sous-jacente (voir l'Annexe B pour plus de détails).

Cette cohérence géométrique se traduit directement par une réduction de la dérive hamiltonienne. Bien que la dérive d'énergie ne soit pas utilisée comme objectif d'optimisation, elle constitue un diagnostic utile de la cohérence numérique et structurelle : les modèles qui violent la géométrie symplectique présentent une injection ou une dissipation systématique d'énergie, alors que le modèle HNN symplectique maintient une dérive faible et bornée sur de longs horizons.

Dérive énergétique et robustesse. La Figure 5 complète l'analyse dans l'espace des phases en présentant la dérive relative de l'Hamiltonien appris H_θ sur de longs horizons de prédiction sous observations bruitées, ainsi que la variabilité entre les différentes graines aléatoires. Comme H_θ définit une jauge d'énergie latente plutôt que l'énergie physique réelle, la dérive rapportée doit être interprétée comme un diagnostic de cohérence interne et non comme une mesure de l'erreur d'énergie physique. Sur l'ensemble des systèmes, la contrainte hamiltonien réduit considérablement la variabilité d'une exécution à l'autre, produisant des bandes de confiance plus étroites et une meilleure robustesse face aux erreurs de prédiction cumulatives. Cet effet est particulièrement marqué sur Hénon–Heiles et Kepler, où les prédictions non contraintes ou non symplectiques présentent une dérive croissant rapidement et une forte dispersion selon les graines, tandis que les modèles hamiltoniens produisent un comportement nettement plus cohérent. Le couplage de la structure hamiltonienne avec une discrétisation symplectique borne davantage la dérive sur tout l'horizon, indiquant une réduction de l'injection ou de la dissipation d'énergie induite par le solveur.

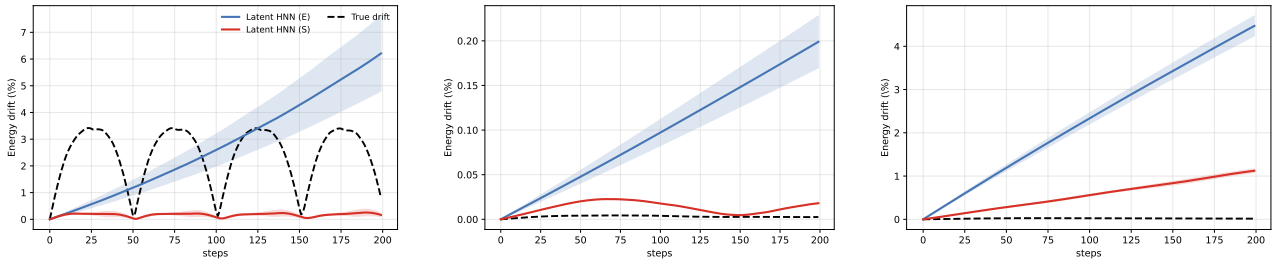
Notamment, l'hamiltonien appris peut présenter une dérive plus faible que l'énergie évaluée sur les trajectoires de référence. Ceci n'est pas contradictoire : cette dernière reflète les effets de discrétisation et la métrique de dérive relative choisie, tandis que le premier correspond à un flot latent structurellement contraint qui impose une évolution conservative par construction. Dans l'ensemble, ces diagnostics renforcent l'interprétation géométrique : la structure hamiltonienne limite l'espace des dynamiques latentes admissibles, et l'intégration symplectique préserve cette structure au cours du temps, garantissant ainsi des prédictions à long horizon plus stables et fiables sous observabilité partielle.

Ablation : Matrice de masse identité vs apprise. En suivant l'interprétation géométrique ci-dessus, nous isolons maintenant le rôle de la métrique latente induite par l'énergie cinétique. Fixer la matrice de masse à l'identité est un choix naturel et physiquement cohérent pour les systèmes considérés, dont les vrais hamiltoniens possèdent une énergie cinétique séparable avec une masse constante. Cependant, dans notre formulation latente, les coordonnées (y, h) sont apprises et ne coïncident pas nécessairement avec les variables physiques canoniques. En conséquence, même lorsque le système sous-jacent admet une matrice de masse constante, la géométrie latente effective induite par l'observabilité partielle et les représentations apprises est généralement non euclidienne.

Empiriquement, contraindre $M_\theta(y) = I$ dégrade significativement les performances à long horizon sur le système de Hénon–Heiles. Comme le montre le Tableau 2, la variante avec une matrice de masse apprise dépendant de l'état atteint une RMSE à long horizon ($L = 200$) nettement plus faible sous observations partielles. Bien que les deux mo-

Mass matrix	**RMSE** ($L = 200$)
Learned $M_\theta(y)$	0.0420 ± 0.0025
Identity $M = I$	0.0597 ± 0.0076

TABLE 2 – Ablation de la matrice de masse latente sur le système de Hénon–Heiles avec observations partielles ($C = W = 50$). Fixer la masse à l'identité dégrade la précision à long horizon malgré la préservation de la structure hamiltonienne.



(a) Pendulum (b) Hénon-Heiles (c) Kepler

FIGURE 5 – Dérive hamiltonienne relative au cours du temps, exprimée en pourcentage de l'énergie initiale. Les lignes pleines représentent la moyenne sur les graines (seeds) et les zones ombrées indiquent l'écart-type ± 1.

dèles restent hamiltoniens et conservent l'énergie en temps continu, fixer la masse limite la puissance expressive de l'hamiltonien latent et restreint la flexibilité du terme cinétique. Géométriquement, cela correspond à imposer une métrique latente globalement euclidienne, ce qui est insuffisant pour capturer la structure courbe de l'espace des phases observée dans les visualisations précédentes.

Cette restriction impacte également l'inférence de l'état initial : en imposant une relation rigide entre les moments latents et la dynamique observée, elle réduit la capacité de l'encodeur à inférer de manière robuste $\hat{h}_0$ à partir d'observations bruitées et partielles. L'apprentissage d'une matrice de masse dépendant de l'état pallie ces limitations en adaptant la métrique latente à la géométrie de l'espace des phases inféré, ce qui se traduit par une meilleure robustesse et une précision accrue à long horizon.

5 Limitations

Notre étude se concentre sur des systèmes hamiltoniens simulés de faible dimension, qui sont bien adaptés pour isoler l'effet de la structure hamiltonienne et de la discrétisation symplectique sous observabilité partielle. Bien que les résultats soulignent systématiquement des gains en termes de robustesse et de stabilité à long horizon, l'extension de cette approche à des systèmes de plus haute dimension et à des modalités d'observation plus complexes (par exemple, des images ou des schémas de détection irréguliers) reste une direction ouverte. Le modèle dynamique proposé cible principalement des régimes approximativement conservatifs et ne capture pas explicitement la dissipation ou les forçages externes. L'extension du cadre à des contextes non conservatifs ou contrôlés, par exemple via des formulations port-hamiltoniennes ou des approches connexes, constitue une avenue naturelle pour de futurs travaux.

Enfin, sous observabilité partielle, l'état latent initial est inféré à partir d'une fenêtre de contexte finie, ce qui induit un plancher d'incertitude irréductible en présence de bruit et peut limiter la précision à long horizon dans des régimes difficiles. Bien que nos résultats montrent que les contraintes géométriques atténuent l'amplification d'erreur, des objectifs d'inférence plus structurés (par exemple, des encodeurs basés sur le filtrage ou le lissage) pourraient encore améliorer la robustesse.

6 Conclusion

Nous avons introduit un cadre de modélisation hamiltonien latent pour l'apprentissage de systèmes dynamiques sous observabilité partielle, dans lequel un paramétrage hamiltonien structuré est combiné à une inférence d'état basée sur le contexte et à une intégration temporelle symplectique.

Plutôt que de viser la récupération exacte de l'énergie physique, l'approche proposée exploite la structure hamiltonienne comme un biais inductif géométrique qui contraint l'espace des flots latents admissibles.

À travers des expériences approfondies sur des systèmes non linéaires canoniques, nous avons montré que ce biais géométrique apporte des bénéfices concrets pour la prédiction à long horizon. En particulier, le paramétrage hamiltonien réduit systématiquement la variance entre les initialisations aléatoires, tandis que les projections symplectiques atténuent la dérive induite par le solveur et la distorsion artificielle de l'espace des phases.

Ensemble, ces effets se traduisent par des prédictions plus stables et plus fiables, en particulier dans les régimes où les erreurs cumulatives et l'observabilité partielle prédominent.

Notre analyse souligne que l'hamiltonien appris doit être interprété comme une jauge d'énergie latente : il définit un système de coordonnées conservatif dans lequel la dynamique évolue de manière cohérente, même lorsque l'énergie physique n'est pas directement observable ou est affectée par la discrétisation et le bruit.

De ce point de vue, la dérive énergétique sert de diagnostic de cohérence interne et de robustesse numérique, plutôt que de mesure de fidélité physique.

Dans l'ensemble, ce travail positionne la structure hamiltonienne et l'intégration symplectique comme des outils complémentaires pour contrôler l'amplification des erreurs à long terme dans les modèles dynamiques latents.

En imposant explicitement des contraintes géométriques à la fois au niveau du modèle et du solveur, notre approche offre une voie rigoureuse et pratique vers une prédiction robuste à long horizon sous observabilité partielle, et pose les bases de l'extension de l'apprentissage préservant la structure à des systèmes dynamiques plus complexes, de plus grande dimension et pilotés par les données.

Références

[1] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, 2018.

[2] Miles D. Cranmer, Sam Greydanus, Stephan Hoyer, Peter W. Battaglia, David N. Spergel, and Shirley Ho. Lagrangian neural networks. *CoRR*, abs/2003.04630, 2020.

[3] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In *Advances in Neural Information Processing Systems*, pages 451–457, 2000.

[4] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, 2019.

[5] Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2555–2565, 2019.

[6] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer, 2006.

[7] Rahul G. Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *AAAI Conference on Artificial Intelligence*, pages 2101–2109, 2017.

[8] Kookjin Lee and Kevin T. Carlberg. Deep conservation : A latent-dynamics model for exact satisfaction of physical conservation laws. In *AAAI Conference on Artificial Intelligence*, 2019.

[9] Fabien Lionti, Nicolas Gutowski, Sébastien Aubin, and Philippe Martinet. Bias-variance analysis of multi-step loss functions for dynamical system identification. In *Proceedings of the International Joint Conference on Neural Networks*, 2025.

[10] Lennart Ljung. *System Identification : Theory for the User*. Prentice Hall PTR, 1999.

[11] Michael Lutter, Christian Ritter, and Jan Peters. Deep lagrangian networks : Using physics as model prior for deep learning. *CoRR*, abs/1907.04490, 2019.

[12] Spencer M. Richards, Felix Berkenkamp, and Andreas Krause. The lyapunov neural network : Adaptive stability certification for safe learning of dynamic systems. *CoRR*, abs/1808.00924, 2018.

[13] Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. *Latent ODEs for Irregularly-Sampled Time Series*. Curran Associates, 2019.

[14] Peter Toth, Danilo Jimenez Rezende, Andrew Jaegle, Sébastien Racanière, Aleksandar Botev, and Irina Higgins. Hamiltonian generative networks. *CoRR*, abs/1909.13789, 2019.

[15] Aiqing Zhu, Pengzhan Jin, Beibei Zhu, and Yifa Tang. On numerical integration in neural ordinary differential equations. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27527–27547, 2022.

A Inférence de l'état initial : l'allongement du contexte améliore le rapport signal/bruit jusqu'à un plancher de variance

Nous fournissons des détails mathématiques supplémentaires étayant la discussion sur la robustesse en §2. La quantité clé gouvernant la robustesse à long horizon sous observabilité partielle est la covariance de l'état latent initialisé, $\Sigma_0 = \mathrm{Cov}(z_0)$. Nous expliquons comment la longueur du contexte C influence Σ_0, pourquoi un plancher de variance irréductible subsiste, et comment des rollouts préservant la structure sur un horizon W contrôlent la propagation de cette incertitude.

Configuration (observations partielles et état initial filtré). Nous considérons un état latent augmenté $z_t = [y_t, h_t]^\top$ ainsi que des observations bruitées

$$y_t^{\mathrm{obs}} = y_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, R), \tag{7}$$

avec une dynamique latente déterministe $z_{t+1} = f_\theta(z_t)$ induite par un champ de vecteurs hamiltonien (intégré numériquement comme dans (4)–(5)). Un encodeur de contexte E_ϕ (GRU) projette une fenêtre d'observations $Y^{\mathrm{obs}}_{t-C:t}$ vers une estimation $\hat{h}_0 = E_\phi(Y^{\mathrm{obs}}_{t-C:t})$, et nous initialisons le rollout de prédictions autorégressives par

$$z_0 = [y_0^{\mathrm{obs}}, \hat{h}_0]^\top. \tag{8}$$

D'après (8), la covariance initiale se décompose en une incertitude portant sur le moment latent inféré $\hat{h}_0$ et une incertitude associée à la coordonnée initiale observée y_0^{obs}. Ces deux contributions jouent des rôles fondamentalement distincts dans la robustesse à long horizon.

Propagation sur W et rôle de la "rigidité" symplectique. Dans [9], la fonction objectif associé à un rollout de prédictions autorégressives à T pas admet la décomposition biais–variance suivante :

$$L(\theta) = \mathrm{Bias}_T^2 + \mathrm{Var}_T \tag{9}$$

avec

$$\mathrm{Var}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathrm{Tr}(\Sigma_t)], \qquad \Sigma_{t+1} = J_f(\mu_t)\Sigma_t J_f(\mu_t)^\top, \tag{10}$$

sous une linéarisation au premier ordre autour de la trajectoire moyenne μ_t. En déroulant (10), on obtient $\Sigma_t = A_t \Sigma_0 A_t^\top$ avec $A_t = \prod_{k=0}^{t-1} J_f(\mu_k)$, ce qui montre que Var_T hérite d'une contribution non nulle provenant de la covariance initiale Σ_0.
Dans le régime hamiltonien considéré, la dynamique en temps continu préserve la structure géométrique et demeure typiquement non contractante et non expansive. Ainsi, l'incertitude initiale représentée par Σ_0 persiste à long terme plutôt que de s'atténuer. L'emploi d'un schéma d'intégration symplectique impose une rigidité numérique aux projections sur W pas : la préservation discrète de la forme symplectique empêche l'apparition d'effets dissipatifs ou instables artificiels, fréquents avec des solveurs non symplectiques. Par suite, accroître W stabilise l'objectif d'entraînement et améliore le rapport signal/bruit, en effectuant une moyenne sur des trajectoires plus longues et géométriquement consistantes.

Effet de la longueur du contexte C sur la qualité d'inférence de $\hat{h}_0$. Nous considérons l'estimation du moment latent initial $\hat{h}_0$ à partir d'une séquence d'observations bruitées $Y^{\mathrm{obs}}_{0:C}$ régie par (7). Soit $\mu_t(h_0)$ la prédiction déterministe de y_t conditionnelle à h_0, obtenue par propagation de la dynamique latente. La log-vraisemblance associée à la fenêtre s'écrit

$$\mathcal{L}(h_0) = \frac{1}{2}\sum_{t=0}^{C} \|y_t^{\mathrm{obs}} - \mu_t(h_0)\|^2_{R^{-1}} + \text{const.} \tag{11}$$

En notant $J_t = \partial \mu_t(h_0)/\partial h_0$, une approximation de type Gauss–Newton, équivalente à l'information de Fisher, donne

$$\mathcal{I}_C(h_0) \approx \sum_{t=0}^{C} J_t^\top R^{-1} J_t, \tag{12}$$

d'où découle la borne de Cramér–Rao

$$\mathrm{Cov}(\hat{h}_0) \succeq \mathcal{I}_C(h_0)^{-1}. \tag{13}$$

Étant donné que l'augmentation de C ajoute des termes définis positifs (semi-définis positifs) dans (12), la matrice d'information $\mathcal{I}_C$ est non décroissante en C. Lorsqu'elle est de rang plein, la borne (13) décroît donc avec C. Ce résultat formalise le rôle de l'encodeur de contexte comme filtre temporel : des fenêtres plus longues fournissent des mesures redondantes qui améliorent le rapport signal/bruit de $\hat{h}_0$. Cependant, l'efficacité de ce filtrage dépend de manière cruciale de la qualité du signal de supervision utilisé durant l'entraînement. Lorsque l'horizon de rollout W est suffisamment grand et que l'intégration est symplectique, l'objectif d'entraînement moyenne les erreurs de prédiction le long de trajectoires longues et géométriquement cohérentes, produisant ainsi un signal d'apprentissage à fort rapport signal/bruit. Dans ce régime, l'encodeur peut pleinement exploiter sa capacité de filtrage à mesure que C augmente. À l'inverse, pour de faibles valeurs de W ou en présence de rollouts non symplectiques, le signal de supervision devient lui-même bruité et biaisé par des artefacts numériques, ce qui limite le gain apporté par l'augmentation de C, malgré la croissance théorique de l'information de Fisher.

Plancher de variance induit par l'ancrage observationnel. Les projections sont initialisées à partir de la coordonnée observée y_0^{obs} comme défini en (8). Ainsi, indépendamment de la qualité d'inférence de $\hat{h}_0$, la covariance $\text{Cov}(y_0^{\text{obs}}) = R$ impose

$$\Sigma_0 \succeq \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix}, \tag{14}$$

ce qui définit une borne inférieure irréductible sur l'incertitude initiale.
En outre, en régime d'observabilité partielle, la matrice d'information de Fisher (12) peut être singulière, empêchant l'identification complète de certaines directions latentes à partir d'un contexte de longueur finie. Il en résulte que l'allongement du contexte C réduit l'incertitude sur $\hat{h}_0$ sans pouvoir annuler l'incertitude globale Σ_0.

Interaction entre la longueur du contexte C et l'horizon de rollout W. L'analyse précédente met en évidence que les bénéfices associés à l'augmentation de la longueur du contexte C et de l'horizon de rollout W sont intrinsèquement couplés. La longueur du contexte C régit l'efficacité statistique de l'inférence de l'état initial en agissant comme un filtre temporel, tandis que l'horizon W détermine le rapport signal/bruit de l'objectif d'entraînement lui-même. Lorsque W est suffisamment grand et que les rollouts sont produits par un intégrateur préservant la structure (symplectique), le signal de supervision devient stable et de faible variance, permettant à l'encodeur d'exploiter pleinement des fenêtres de contexte plus longues. Dans ce régime, accroître C réduit principalement la variance du moment latent inféré $\hat{h}_0$, ce qui conduit à des prédictions à long horizon plus cohérentes d'une graine aléatoire à l'autre.
À l'inverse, lorsque W est trop court ou lorsque les rollouts sont affectés par des artefacts numériques, le signal d'entraînement devient bruité et biaisé, et la capacité de filtrage de l'encodeur sature prématurément : l'augmentation de C produit alors des gains décroissants malgré l'accroissement théorique de l'information. Enfin, indépendamment du choix de (C, W), un plancher de variance irréductible subsiste sous observabilité partielle, en raison de l'ancrage dans l'observation initiale bruitée y_0^{obs}. Ce mécanisme explique les compromis empiriques observés dans nos expériences et motive la sélection conjointe de (C, W), ainsi que l'intégration symplectique, comme un problème de co-conception pour un apprentissage robuste à long horizon.

B Sur la liberté de jauge de l'Hamiltonien appris

Dans ce travail, l'Hamiltonien H_θ est appris dans un système de coordonnées latent et ne doit pas être interprété comme l'énergie physique réelle du système sous-jacent. Il définit plutôt une fonction d'énergie latente, identifiable uniquement à une transformation de jauge près.

Liberté de jauge et non-identifiabilité de l'Hamiltonien appris. Un même flot hamiltonien peut être représenté par différentes fonctions d'énergie, liées par des transformations de coordonnées. Les transformations affines $H_\theta \mapsto aH_\theta + b$ $(a > 0)$ préservent la structure du champ hamiltonien à une dilatation temporelle près, tandis que des transformations dépendantes de l'état peuvent être interprétées comme des modifications de la métrique induite par l'énergie cinétique. Il en résulte que ni l'échelle absolue ni l'offset de H_θ ne sont identifiables de manière unique à partir des seules observations.

Implications pour les visualisations dans l'espace des phases. Dans la mesure où le terme cinétique $h^\top M_\theta(y)^{-1} h$ induit une métrique latente sur l'espace des phases, différents choix de jauge (par exemple via des matrices de masse apprises distinctes) entraînent des redimensionnements géométriques différents des trajectoires. Cela explique pourquoi la taille et la forme des orbites dans l'espace des phases peuvent varier d'un modèle à l'autre, même lorsque la structure qualitative de la dynamique est préservée. Il est essentiel de noter que ces redimensionnements n'affectent ni la symplecticité ni l'organisation topologique des trajectoires, qui constituent les objets principaux de notre analyse.

Rôle de la dérive de l'Hamiltonien appris. Compte tenu de la liberté de jauge, le niveau absolu de H_θ ne possède pas de signification physique directe. La dérive relative de H_θ est donc interprétée comme un indicateur de cohérence interne plutôt que comme une mesure d'erreur d'énergie physique. Une dérive croissante traduit typiquement des incohérences entre la dynamique apprise et sa discrétisation (intégration non symplectique, distorsions géométriques), alors qu'une dérive faible et bornée indique une préservation effective de la structure. Cette lecture rejoint l'analyse biais–variance de l'Annexe A, dans laquelle les violations structurelles induisent une amplification numérique de l'incertitude.

C Systèmes dynamiques et distributions des conditions initiales

Dans cet appendice, nous détaillons les équations dynamiques employées pour la génération des trajectoires, ainsi que les distributions de conditions initiales (CI) propres à chaque système. Les simulations sont réalisées avec un pas de temps constant $\Delta t = 0.02$ et une durée totale de $T = 500$ pas. Sauf mention explicite, les distributions de CI sont identiques pour les phases d'entraînement, de validation et d'évaluation.

Oscillateur de Duffing non amorti (2D). Le système étudié est l'oscillateur de Duffing conservatif, défini par $x = (q, v) \in \mathbb{R}^2$ et régi par

$$\dot{q} = v, \qquad \dot{v} = -\frac{k}{m}\, q - \frac{\alpha}{m}\, q^3. \tag{15}$$

Les paramètres sont fixés à $m = 1$, $k = 1$ et $\alpha = 1$. Les conditions initiales (q_0, v_0) sont tirées indépendamment selon

$$(q_0, v_0) \sim \mathcal{U}([-\beta, \beta]^2), \qquad \beta = 0.5. \tag{16}$$

Pendule simple conservatif (2D). Le système étudié est un pendule sans dissipation, défini par $x = (\theta, \omega) \in \mathbb{R}^2$ et régi par

$$\dot{\theta} = \omega, \qquad \dot{\omega} = -\frac{g}{\ell} \sin(\theta). \tag{17}$$

Les paramètres physiques sont fixés à $g = 9.81$ et $\ell = 1$. Les conditions initiales (θ_0, ω_0) sont tirées indépendamment selon

$$(\theta_0, \omega_0) \sim \mathcal{U}([-\beta, \beta]^2), \qquad \beta = 0.5. \tag{18}$$

Problème de Kepler (4D). Nous considérons le problème de Kepler planaire, défini par les coordonnées canoniques $x = (q_1, q_2, p_1, p_2) \in \mathbb{R}^4$ et l'Hamiltonien

$$H(q, p) = \frac{1}{2}\|p\|_2^2 - \frac{\mu}{\|q\|_2}, \qquad q = (q_1, q_2),\ p = (p_1, p_2), \tag{19}$$

qui induit la dynamique

$$\dot{q} = p, \qquad \dot{p} = -\mu \frac{q}{\|q\|_2^3}. \tag{20}$$

Nous utilisons $\mu = 1$. Les conditions initiales sont générées en échantillonnant un rayon $r \sim \mathcal{U}[r_{\min}, r_{\max}]$ avec $(r_{\min}, r_{\max}) = (0.7, 1.3)$, ainsi qu'un angle $\varphi \sim \mathcal{U}[0, 2\pi]$, puis en posant

$$q_0 = (r\cos\varphi,\ r\sin\varphi), \qquad p_0 = v_{\text{circ}}\,(-\sin\varphi,\ \cos\varphi), \qquad v_{\text{circ}} = \sqrt{\mu/r}. \tag{21}$$

Cette procédure génère des orbites quasi elliptiques avec une vitesse initiale tangentielle.

Système de Hénon–Heiles (4D). Nous considérons le système hamiltonien de Hénon–Heiles, défini par les coordonnées canoniques $x = (q_1, q_2, p_1, p_2) \in \mathbb{R}^4$ et l'Hamiltonien

$$H(q, p) = \frac{1}{2}(p_1^2 + p_2^2) + V(q_1, q_2), \qquad V(q_1, q_2) = \frac{1}{2}(q_1^2 + q_2^2) + \left(q_1^2 q_2 - \frac{1}{3} q_2^3\right). \tag{22}$$

Les équations du mouvement induites sont

$$\dot{q}_1 = p_1, \quad \dot{q}_2 = p_2, \qquad \dot{p}_1 = -\frac{\partial V}{\partial q_1} = -(q_1 + 2q_1 q_2), \qquad \dot{p}_2 = -\frac{\partial V}{\partial q_2} = -(q_2 + q_1^2 - q_2^2). \tag{23}$$

Les conditions initiales sont échantillonnées à une énergie cible fixée E (nous utilisons $E = 0.05$). Nous tirons d'abord $q_0 \sim \mathcal{U}([-\beta, \beta]^2)$ avec $\beta = 0.5$, et nous acceptons les échantillons vérifiant $V(q_0) < E$. Pour chaque q_0 accepté, nous tirons un angle aléatoire $\psi \sim \mathcal{U}[0, 2\pi]$ et ajustons la norme du moment de façon à correspondre à l'énergie cinétique résiduelle :

$$\|p_0\|_2 = \sqrt{2(E - V(q_0))}, \qquad p_0 = \|p_0\|_2(\cos\psi, \sin\psi). \tag{24}$$

D Architecture du modèle et réglages d'hyperparamètres

Cet appendice résume les architectures et les hyperparamètres par défaut utilisés pour l'ensemble des modèles présentés en §2 et §3.

Configuration commune (dimensions observées et latentes). Soit $y_t \in \mathbb{R}^{d_y}$ les variables observées (configurations) et $h_t \in \mathbb{R}^{d_h}$ l'augmentation latente (interprétée comme des moments dans les modèles hamiltoniens latents). Dans le cadre partiellement observé, nous utilisons $d_h = d_y$ afin de permettre un appariement symplectique canonique. Dans le cadre pleinement observé, nous posons $d_h = 0$ et y_t contient l'état canonique complet (par exemple $[q, p]$).

Encodeur de contexte (GRU). L'inférence de l'état latent initial repose, pour tous les modèles, sur un encodeur commun E_ϕ :

- **Backbone :** un réseau récurrent de type GRU, recevant des entrées de dimension d_y, avec une dimension cachée $H_{\text{enc}} = 64$ et $L_{\text{enc}} = 2$ couches empilées.
- **Readout :** l'état caché final est transformé par une projection linéaire vers $\mathbb{R}^{d_h}$ afin d'obtenir l'estimation du moment latent initial $\hat{h}_0$.

Latent NODE (E) : champ de vecteurs augmenté non contraint. La baseline Latent NODE paramètre un champ de vecteurs en temps continu non contraint sur l'état augmenté $z = [y, h] \in \mathbb{R}^{d_y+d_h}$:

$$\dot{z} = f_\theta(z), \tag{25}$$

où f_θ est un perceptron multicouche (MLP) avec activations $\tanh$:

- **Entrée/sortie :** $(d_y + d_h) \to (d_y + d_h)$.
- **Largeur/profondeur :** largeur cachée $H = 64$ et profondeur $D = 3$ (deux couches cachées avec activations $\tanh$, suivies d'une couche de sortie linéaire).
- **Initialisation :** la couche linéaire finale est initialisée à zéro (poids et biais), de sorte que la dynamique initiale soit quasi nulle et que l'entraînement débute à partir de rollouts stables.

Les rollouts sont intégrés à l'aide d'un schéma d'Euler explicite (E).

Latent HNN (E/S) : dynamique augmentée hamiltonienne structurée. Les modèles Latent HNN reposent sur une paramétrisation hamiltonienne de la forme

$$H_\theta(y, h) = \frac{1}{2} h^\top M_\theta(y)^{-1} h + V_\theta(y), \tag{26}$$

où $V_\theta : \mathbb{R}^{d_y} \to \mathbb{R}$ désigne l'énergie potentielle, et où $M_\theta(y)^{-1}$ est une matrice symétrique définie positive (SPD) dépendant de y. Cette paramétrisation est implémentée comme suit :

- **Réseau de potentiel :** $V_\theta(y)$ est modélisé par un MLP $d_y \to 1$, de largeur $H = 64$, de profondeur $D = 3$, avec activations $\tanh$. La dernière couche est initialisée avec de petits poids aléatoires (gaussiens, écart-type 10^{-3}) et un biais nul.
- **Réseau de masse :** $M_\theta(y)^{-1}$ est paramétrée via un facteur de Cholesky $L_\theta(y)$. Les coefficients triangulaires inférieurs de $L_\theta(y) \in \mathbb{R}^{d_h \times d_h}$ sont prédits par un MLP $d_y \to \frac{d_h(d_h+1)}{2}$, de largeur $H = 64$, profondeur $D = 3$, avec activations $\tanh$; la couche de sortie est initialisée à zéro. Les éléments diagonaux de $L_\theta(y)$ sont passés à travers une transformation softplus et décalés par ε afin de garantir une positivité stricte.
- **Construction SPD :** nous définissons

$$M_\theta(y)^{-1} = L_\theta(y) L_\theta(y)^\top + \varepsilon I, \qquad \varepsilon = 10^{-6}. \tag{27}$$

Discrétisation temporelle. Deux schémas d'intégration sont considérés :

- **Euler explicite (E),** défini par $z_{k+1} = z_k + \Delta t\, \dot{z}_k$.
- **Euler symplectique « kick–drift » (S),** employé pour les dynamiques hamiltoniennes et cohérent avec (4)–(5). Sous observabilité complète ($y = [q, p]$), la mise à jour est appliquée après séparation des variables canoniques (q, p).

E Détails des experiences

TABLE 3 – Balayage détaillé des fenêtres sur le système de Duffing : RMSE de test (moyenne ± écart-type sur les graines aléatoires) pour différentes valeurs de (C, W) et pour chaque méthode. Les valeurs en gras indiquent le meilleur couple (C, W) (RMSE moyenne minimale) pour chaque ligne.

Duffing — Observations propres

Model	$C = 10$			$C = 25$			$C = 50$		
	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$
Latent NODE (E)	0.2920 ± 0.1727	0.1794 ± 0.1665	0.1540 ± 0.1428	0.0302 ± 0.0051	0.0291 ± 0.0044	0.0267 ± 0.0039	0.0562 ± 0.0147	0.0389 ± 0.0060	**0.0238 ± 0.0022**
Latent HNN (E)	0.0465 ± 0.0129	0.0212 ± 0.0203	0.0168 ± 0.0075	0.0359 ± 0.0178	0.0188 ± 0.0117	0.0173 ± 0.0038	0.0134 ± 0.0039	0.0129 ± 0.0054	**0.0071 ± 0.0005**
Latent HNN (S)	0.0470 ± 0.0201	0.0211 ± 0.0180	0.0165 ± 0.0060	0.0326 ± 0.0246	0.0212 ± 0.0078	0.0171 ± 0.0048	0.0143 ± 0.0071	0.0128 ± 0.0078	**0.0035 ± 0.0011**

Duffing — Observations bruitées

Model	$C = 10$			$C = 25$			$C = 50$		
	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$
Latent NODE (E)	0.1514 ± 0.0002	0.1413 ± 0.0032	0.0806 ± 0.0038	0.1157 ± 0.0204	0.0442 ± 0.0051	**0.0354 ± 0.0023**	0.2433 ± 0.0811	0.0608 ± 0.0142	0.0398 ± 0.0057
Latent HNN (E)	0.1171 ± 0.0170	0.0971 ± 0.0062	0.1263 ± 0.0042	0.1331 ± 0.0663	0.0481 ± 0.0042	0.0365 ± 0.0033	0.1050 ± 0.0798	0.0418 ± 0.0044	**0.0295 ± 0.0030**
Latent HNN (S)	0.1237 ± 0.0349	0.0950 ± 0.0063	0.1261 ± 0.0031	0.1499 ± 0.0856	0.0452 ± 0.0012	0.0355 ± 0.0025	0.1244 ± 0.1061	0.0401 ± 0.0007	**0.0292 ± 0.0032**

TABLE 4 – Balayage détaillé des fenêtres sur le système de Hénon : RMSE de test (moyenne ± écart-type sur les graines aléatoires) pour différentes valeurs de (C, W) et pour chaque méthode. Les valeurs en gras indiquent le meilleur couple (C, W) (RMSE moyenne minimale) pour chaque ligne.

Hénon — Observations propres

Model	$C = 10$			$C = 25$			$C = 50$		
	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$
Latent NODE (E)	0.2049 ± 0.0201	0.1807 ± 0.0234	0.1590 ± 0.0204	0.1847 ± 0.0030	0.1714 ± 0.0002	0.0994 ± 0.0548	0.1895 ± 0.0087	0.1350 ± 0.0369	**0.0381 ± 0.0049**
Latent HNN (E)	0.0870 ± 0.0029	0.0526 ± 0.0041	0.0450 ± 0.0056	0.0584 ± 0.0091	0.0455 ± 0.0048	0.0424 ± 0.0052	0.0685 ± 0.0164	0.0437 ± 0.0024	**0.0367 ± 0.0060**
Latent HNN (S)	0.1196 ± 0.0215	0.0559 ± 0.0067	0.0425 ± 0.0053	0.0743 ± 0.0130	0.0439 ± 0.0047	0.0409 ± 0.0044	0.0805 ± 0.0237	0.0428 ± 0.0030	**0.0379 ± 0.0035**

Hénon — Observations bruitées

Model	$C = 10$			$C = 25$			$C = 50$		
	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$
Latent NODE (E)	0.1987 ± 0.0060	0.1680 ± 0.0007	0.1439 ± 0.0134	0.1887 ± 0.0153	0.1475 ± 0.0206	0.1334 ± 0.0286	0.1814 ± 0.0073	0.1739 ± 0.0019	**0.1124 ± 0.0344**
Latent HNN (E)	0.1614 ± 0.0210	0.0818 ± 0.0072	0.1336 ± 0.0005	0.1416 ± 0.0349	0.0707 ± 0.0058	0.0585 ± 0.0021	0.1280 ± 0.0620	0.0697 ± 0.0026	**0.0568 ± 0.0045**
Latent HNN (S)	0.1526 ± 0.0000	0.0779 ± 0.0000	0.1340 ± 0.0000	0.1586 ± 0.0000	0.0672 ± 0.0000	0.0577 ± 0.0000	0.1921 ± 0.0000	0.0656 ± 0.0000	**0.0524 ± 0.0000**

TABLE 5 – Balayage détaillé des fenêtres sur le système de Kepler : RMSE de test (moyenne ± écart-type sur les graines aléatoires) pour différentes valeurs de (C, W) et pour chaque méthode. Les valeurs en gras indiquent le meilleur couple (C, W) (RMSE moyenne minimale) pour chaque ligne.

Kepler — Observations propres)

Model	$C = 10$			$C = 25$			$C = 50$		
	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$
Latent NODE (E)	0.3049 ± 0.2995	0.2618 ± 0.2547	**0.1884 ± 0.1707**	0.5820 ± 0.0085	0.4653 ± 0.0265	0.3631 ± 0.0230	0.6166 ± 0.0202	0.4879 ± 0.0581	0.3630 ± 0.0361
Latent HNN (E)	0.2499 ± 0.0755	0.0991 ± 0.0164	0.0604 ± 0.0041	0.1555 ± 0.0319	0.0792 ± 0.0110	0.0481 ± 0.0044	0.1561 ± 0.0285	0.0724 ± 0.0125	**0.0399 ± 0.0060**
Latent HNN (S)	0.1434 ± 0.0319	0.0645 ± 0.0112	0.0333 ± 0.0036	0.1125 ± 0.0199	0.0549 ± 0.0075	**0.0324 ± 0.0021**	0.1197 ± 0.0141	0.0613 ± 0.0077	0.0419 ± 0.0024

Kepler — Observations bruitées

Model	$C = 10$			$C = 25$			$C = 50$		
	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$
Latent NODE (E)	0.5872 ± 0.2940	0.3991 ± 0.0160	**0.3494 ± 0.0322**	0.5023 ± 0.2027	0.3875 ± 0.0109	0.3562 ± 0.0270	0.6118 ± 0.3321	0.3944 ± 0.0084	0.3514 ± 0.0445
Latent HNN (E)	0.3518 ± 0.1097	0.0748 ± 0.0131	0.0619 ± 0.0122	0.3582 ± 0.1096	0.0735 ± 0.0100	0.0519 ± 0.0059	0.3040 ± 0.1007	0.0800 ± 0.0031	**0.0516 ± 0.0005**
Latent HNN (S)	0.3690 ± 0.0957	0.1122 ± 0.0061	**0.0523 ± 0.0132**	0.3496 ± 0.1209	0.1230 ± 0.0083	0.0547 ± 0.0059	0.3077 ± 0.0597	0.1318 ± 0.0001	0.0692 ± 0.0000

TABLE 6 – Balayage détaillé des fenêtres sur le système Pendule : RMSE de test (moyenne ± écart-type sur les graines aléatoires) pour différentes valeurs de (C, W) et pour chaque méthode. Les valeurs en gras indiquent le meilleur couple (C, W) (RMSE moyenne minimale) pour chaque ligne.

Pendule — Observations propres

Model	$C = 10$			$C = 25$			$C = 50$		
	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$
Latent NODE (E)	0.1301 ± 0.1059	0.1178 ± 0.0979	0.1082 ± 0.1002	0.0505 ± 0.0034	0.0287 ± 0.0007	**0.0109 ± 0.0035**	0.0495 ± 0.0051	0.0303 ± 0.0017	0.0147 ± 0.0032
Latent HNN (E)	0.0842 ± 0.0455	0.0589 ± 0.0045	0.0477 ± 0.0028	0.0619 ± 0.0070	0.0512 ± 0.0058	**0.0445 ± 0.0033**	0.0674 ± 0.0105	0.0539 ± 0.0054	0.0510 ± 0.0042
Latent HNN (S)	0.0286 ± 0.0206	0.0155 ± 0.0001	0.0037 ± 0.0004	0.0128 ± 0.0021	0.0062 ± 0.0015	**0.0035 ± 0.0005**	0.0132 ± 0.0020	0.0065 ± 0.0003	0.0040 ± 0.0001

Pendule — Observations bruitées

Model	$C = 10$			$C = 25$			$C = 50$		
	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$	$W = 10$	$W = 25$	$W = 50$
Latent NODE (E)	0.0628 ± 0.0177	0.0408 ± 0.0029	0.0342 ± 0.0018	0.0856 ± 0.0118	0.0367 ± 0.0024	**0.0227 ± 0.0014**	0.1047 ± 0.0118	0.0396 ± 0.0009	0.0251 ± 0.0028
Latent HNN (E)	0.0774 ± 0.0049	0.0886 ± 0.0000	0.0578 ± 0.0008	0.0808 ± 0.0088	0.0509 ± 0.0036	**0.0483 ± 0.0018**	0.1000 ± 0.0135	0.0600 ± 0.0072	0.0554 ± 0.0069
Latent HNN (S)	0.0501 ± 0.0168	0.0552 ± 0.0011	0.0295 ± 0.0010	0.0589 ± 0.0097	0.0280 ± 0.0034	0.0166 ± 0.0002	0.0455 ± 0.0041	0.0293 ± 0.0005	**0.0158 ± 0.0010**

Caractérisation de la complémentarité des détecteurs d'anomalies par l'analyse des contributions SHAP

Jordan Levy[1,2], Paul Saves[1], Moncef Garouani[1], Nicolas Verstaevel[1], Benoit Gaudou[1]

[1] IRIT, Université Toulouse Capitole
[2] TwinswHeel, Soben

jordan.levy@irit.fr, paul.saves@irit.fr, moncef.garouani@irit.fr, nicolas.verstaevel@irit.fr, benoit.gaudou@irit.fr

Résumé

La détection d'anomalies non supervisée est un problème difficile en raison de la diversité des distributions de données et de l'absence d'étiquettes. Les méthodes ensemblistes sont souvent adoptées pour pallier ces difficultés en combinant plusieurs détecteurs d'anomalies pour réduire les biais individuels et augmenter la robustesse. Cependant, construire un ensemble véritablement complémentaire reste difficile car de nombreux détecteurs reposent sur des critères de discrimination similaires et finissent par produire des scores d'anomalie redondants. Par conséquent, le potentiel de l'apprentissage ensembliste est souvent limité par la difficulté d'identifier des modèles qui capturent vraiment différents types d'irrégularités. Pour remédier à cela, nous proposons une méthodologie pour caractériser les détecteurs d'anomalies à travers leurs mécanismes de décision. En utilisant les explications additives de Shapley (SHAP), nous quantifions comment chaque modèle attribue de l'importance aux caractéristiques d'entrée, et nous utilisons ces profils d'attribution pour mesurer la similarité entre les détecteurs. Nous montrons que les détecteurs ayant des explications similaires ont tendance à produire des scores d'anomalie corrélés et à identifier des anomalies qui se chevauchent largement. Inversement, la divergence des explications indique de manière fiable un comportement de détection complémentaire. Nos résultats démontrent que les métriques basées sur les explications offrent un critère différent, souvent meilleur, des sorties brutes pour sélectionner des modèles dans un ensemble. Cependant, nous démontrons également que la diversité seule est insuffisante ; une performance individuelle élevée des détecteurs d'anomalies reste un prérequis pour des ensembles efficaces. En ciblant explicitement la diversité des explications tout en maintenant la qualité des modèles, nous sommes capables de construire des ensembles plus diversifiés, plus complémentaires et finalement plus efficaces pour la détection d'anomalies non supervisée.

Mots-clés

Détection d'anomalies non supervisée, Modèle ensembliste, Sélection de modèles, Explicabilité.

Abstract

Unsupervised anomaly detection is a challenging problem due to the diversity of data distributions and the lack of labels. Ensemble methods are often adopted to mitigate these challenges by combining multiple anomaly detectors, which can reduce individual biases and increase robustness. Yet building an ensemble that is genuinely complementary remains challenging, since many detectors rely on similar decision cues and end up producing redundant anomaly scores. As a result, the potential of ensemble learning is often limited by the difficulty of identifying models that truly capture different types of irregularities. To address this, we propose a methodology for characterizing anomaly detectors through their decision mechanisms. Using SHapley Additive exPlanations, we quantify how each model attributes importance to input features, and we use these attribution profiles to measure similarity between detectors. We show that detectors with similar explanations tend to produce correlated anomaly scores and identify largely overlapping anomalies. Conversely, explanation divergence reliably indicates complementary detection behavior. Our results demonstrate that explanation-driven metrics offer a different, usually better, criterion than raw outputs for selecting models in an ensemble. However, we also demonstrate that diversity alone is insufficient ; high individual model performance remains a prerequisite for effective ensembles. By explicitly targeting explanation diversity while maintaining model quality, we are able to construct ensembles that are more diverse, more complementary, and ultimately more effective for unsupervised anomaly detection.

Keywords

Unsupervised Anomaly Detection, Ensemble Learning, Model Selection, Explainable AI.

1 Introduction

La détection d'anomalies est un problème difficile, principalement en raison de la nature intrinsèque des anomalies. Les anomalies sont définies comme des déviations par rapport à ce qui est considéré comme un comportement normal [5]. Par conséquent, selon la façon dont cette normalité est définie, un algorithme ou un autre peut être plus adapté pour une

bonne détection. Par exemple, certaines méthodes peuvent définir la normalité en termes géométriques et utiliser des règles basées sur la distance pour repérer les valeurs aberrantes, tandis que d'autres utilisent des hypothèses probabilistes et signalent les instances rares ou à faible probabilité comme des anomalies [19].

Dans de nombreux contextes réels, l'obtention d'anomalies étiquetées est coûteuse ou irréalisable car les événements anormaux sont rares, coûteux à produire ou dangereux à provoquer (*e.g.*, dans l'industrie nucléaire la détection doit fonctionner sans attendre que des défauts se produisent [17]). Par conséquent, les approches semi-supervisées et non supervisées sont souvent préférées. Dans la détection d'anomalies non supervisée (Unsupervised Anomaly Detection, UAD), les spécialistes, ne pouvant pas définir exhaustivement le comportement "normal" à la main, s'appuient généralement sur des méthodes d'apprentissage automatique. Ainsi, au lieu de définir eux-mêmes la normalité, ils utilisent les hypothèses sous-jacentes des algorithmes (géométriques, probabilistes, etc.). Cependant, aucune hypothèse n'est garantie d'être appropriée pour toutes les applications [1], et comme indiqué dans le théorème du "*no free lunch*", aucun détecteur unique ne surpasse systématiquement les autres sur tous les jeux de données [22, 9]. Néanmoins, le choix de l'algorithme reste critique, car une sélection inappropriée peut conduire à de mauvaises performances [1]

Une stratégie courante pour atténuer ce problème consiste à combiner plusieurs détecteurs au sein d'une méthode ensembliste, en tirant parti des atouts de chaque détecteur et de la diversité de leurs hypothèses sur ce qui constitue un comportement normal. L'apprentissage ensembliste a été largement adopté pour cette raison et a démontré de solides performances empiriques dans divers scénarios de détection d'anomalies [1]. En agrégeant des détecteurs avec divers biais inductifs, les approches ensemblistes peuvent capturer une plus grande variété de types d'anomalies. Cette diversité se traduit souvent par une meilleure couverture de détection [21].

Un défi clé dans l'apprentissage ensembliste est la sélection de détecteurs appropriés. Bien qu'une approche ensembliste robuste nécessite théoriquement des détecteurs à la fois diversifiés et performants [21], l'identification de tels comportements complémentaires reste un problème ouvert [13].

Pour relever ce défi, nous proposons une nouvelle méthodologie qui caractérise le comportement des détecteurs d'anomalies en utilisant les explications additives de Shapley (SHAP) [12]. Contrairement aux approches reposant uniquement sur les sorties, nous nous concentrons sur la compréhension des mécanismes de décision internes des détecteurs d'anomalies. Notre étude révèle que les détecteurs ayant des schémas d'explication similaires ont tendance à produire des scores d'anomalie redondants, tandis que la divergence des explications est un indicateur fort de complémentarité. Par conséquent, nous démontrons que la sélection de détecteurs basée sur leur comportement d'explication conduit à une amélioration des performances du modèle ensembliste. Cela conduit à trois contributions majeures :

— Une analyse des algorithmes d'UAD basée sur leurs explications SHAP, qui démontre que leurs comportements sont corrélés à la similarité de leurs sorties.
— Une comparaison entre les explications SHAP et les similarités des sorties de modèles pour sélectionner des modèles de détection d'anomalies diversifiés.
— Une étude empirique quantifiant l'importance relative de la diversité par rapport à la performance individuelle, démontrant que la qualité du modèle reste un prérequis critique.

Cet article est structuré comme suit : la Section 2 présente les travaux connexes. La méthodologie de notre approche est décrite dans la Section 3. Les résultats expérimentaux sont présentés en Section 4. Enfin, nous concluons l'article dans la Section 5.

2 Travaux Connexes

2.1 Algorithmes de détection d'anomalies non supervisée

L'UAD est un problème largement étudié [19] avec un grand nombre d'algorithmes, chacun fondé sur des hypothèses différentes concernant la nature des anomalies. Des bibliothèques comme PyOD [23] ont été introduites pour standardiser l'utilisation de ces modèles. Parmi ceux-ci, les méthodes basées sur la distance (KNN, LOF, CBLOF) caractérisent les anomalies en fonction de leur éloignement des points voisins. De leur côté, les algorithmes basés sur la reconstruction (AutoEncoder, PCA) détectent les anomalies en apprenant à reconstruire des modèles de données normaux. D'autres approches, telles que HBOS, ECOD et COPOD, exploitent des hypothèses probabilistes, tandis que OCSVM et DeepSVDD utilisent la classification à une classe. Enfin, des algorithmes comme Isolation Forest et LODA reposent respectivement sur le partitionnement aléatoire et la projection spatiale.

2.2 Sélection de modèles pour la détection d'anomalies non supervisée

La sélection d'un modèle approprié pour une tâche d'UAD est communément appelée Sélection de Modèle de Valeurs Aberrantes Non Supervisée (Unsupervised Outlier Model Selection, UOMS). L'UOMS est un problème particulièrement difficile qui a gagné en importance à mesure que le nombre de modèles disponibles, comprenant diverses familles algorithmiques et configurations d'hyperparamètres, continue de s'étendre avec l'apparition de nouvelles méthodes dans la littérature [19].

Une stratégie courante en UOMS consiste à estimer la qualité du modèle directement à partir de données non étiquetées de manière non supervisée, pour sélectionner les meilleurs modèles. Ces approches se divisent en deux catégories principales. Les méthodes *autonomes* (stand-alone) [8, 16] calculent un score non supervisé pour chaque détecteur indépendamment, tandis que les méthodes *basées sur le consensus* (consensus-based) évaluent les détecteurs en mesurant l'accord au sein d'un groupe de modèles et en sélectionnant ceux qui se conforment le mieux au groupe [6, 11]. Les preuves empiriques indiquent que les critères d'évaluation

autonomes échouent souvent à fournir des résultats cohérents ou fiables en pratique. En revanche, les approches basées sur le consensus, malgré leur coût de calcul plus élevé, sont apparues comme une stratégie plus efficace et prometteuse pour l'UOMS [13].

Il existe peu de travaux dans la littérature scientifique sur l'UOMS pour l'apprentissage d'ensemble. Dans [21], les auteurs montrent que la diversité des hypothèses algorithmiques tend à donner de meilleurs résultats. Pour caractériser la diversité, les auteurs utilisent une corrélation de Pearson pondérée entre les scores d'anomalie des modèles. Cependant, ils soulignent également que la diversité est importante, mais que les algorithmes choisis doivent déjà avoir de bonnes performances sur le jeu de données pour obtenir un modèle d'ensemble performant. À partir de l'hypothèse précédente, les auteurs de [18] ont introduit SELECT, qui, au lieu d'essayer de sélectionner des modèles en fonction de leur diversité, sélectionne des modèles en fonction de leurs performances à partir d'une pseudo-étiquette de vérité terrain créée à partir des scores d'anomalie. Plus récemment, dans [3], les auteurs font de la détection d'anomalies dans les séries temporelles en utilisant un ensemble de plusieurs auto-encodeurs convolutionnels. Ils optimisent la diversité en intégrant une métrique basée sur la dissimilarité des sorties de reconstruction. Ils démontrent également que cette stratégie axée sur la diversité améliore les performances de détection.

Bien qu'il ait été démontré que les sorties des modèles comme les scores d'anomalies peuvent refléter la diversité, dans ce travail, nous étudions une méthode de consensus basée sur les valeurs SHAP, qui caractérisent le comportement des modèles de détection d'anomalies sur différents jeux de données.

2.3 Explicabilité

L'Intelligence Artificielle Explicable (Explainable Artificial Intelligence, XAI) cherche à rendre les modèles complexes transparents en produisant des représentations interprétables par l'homme sur la manière dont les entrées, la structure du modèle et l'incertitude produisent des sorties particulières [14]. Parmi les méthodes d'explication locales et agnostiques au modèle, les explications additives de Shapley (SHAP) sont largement utilisées car elles fournissent des attributions de caractéristiques axiomatiques au niveau de l'instance [12].

Au-delà de l'interprétabilité, les représentations SHAP ont été exploitées pour comparer des modèles et former des représentations informatives : des travaux récents utilisent des attributions SHAP agrégées pour identifier des familles de modèles ou pour servir de caractéristiques pour des tâches d'apprentissage en aval [20, 7]. À notre connaissance, l'exploitation de SHAP pour la comparaison et la sélection de modèles n'a pas été systématiquement étudiée dans le cadre de l'UAD, où l'absence d'étiquettes rend la sélection des modèles particulièrement difficile.

3 Méthodologie

3.1 Cadre du problème

Nous considérons le problème d'UAD, où l'objectif est d'identifier des échantillons anormaux au sein d'un jeu de données $\mathcal{D} = \{x_1, \ldots, x_n\}$ contenant n instances dans un espace de caractéristiques à d dimensions. Soit $\mathcal{M}$ un ensemble de m modèles de détection d'anomalies $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$. Pour un jeu de données donné, chaque modèle M_i produit un vecteur de score d'anomalie $s_i = M_i(\mathcal{D}) \in \mathbb{R}^n$ où $s_i^{(k)}$ est le score d'anomalie du $k^{\text{ième}}$ point d'entrée selon le $i^{\text{ième}}$ modèle. Chaque modèle produit également un vecteur de prédictions binaires $a_i \in \{0,1\}^n$, obtenu par un seuillage du vecteur de scores d'anomalie, indiquant directement la présence ou l'absence d'anomalies. Plus précisément pour chaque instance k, la prédiction est $a_i^{(k)} = 0$ si $s_i^{(k)} < \tau_i$, et 1 sinon. Le paramètre τ_i est un seuil de décision interne, propre à chaque modèle i. Notre objectif est d'analyser le comportement de ces modèles et d'identifier des groupes de modèles qui partagent des structures interprétatives similaires.

3.2 Similarité des modèles à partir des explications

Pour chaque modèle M_i, nous calculons sa matrice d'explication SHAP $Sh_i \in \mathbb{R}^{n \times d}$, où $Sh_i^{(k)} \in \mathbb{R}^d$ est le vecteur représentant la contribution de chaque caractéristique au score d'anomalie de l'instance x_k.
Nous définissons la similarité de comportement entre deux modèles M_i et M_j comme la corrélation de Pearson moyenne par instance entre leurs vecteurs SHAP :

$$\rho_{ij}^{\text{PS}} = \frac{1}{n} \sum_{k=1}^{n} \text{corr}\big(Sh_i^{(k)}, Sh_j^{(k)}\big).$$

Pour capturer la cohérence du classement entre les importances des caractéristiques plutôt que les magnitudes brutes, nous calculons également une similarité basée sur le Gain Cumulé Actualisé Normalisé (Normalized Discounted Cumulative Gain, NDCG) :

$$\rho_{ij}^{\text{NDCG}} = \frac{1}{2n} \sum_{k=1}^{n} \Big(\text{NDCG}\left(|Sh_i^{(k)}|, |Sh_j^{(k)}|\right) + \text{NDCG}\left(|Sh_j^{(k)}|, |Sh_i^{(k)}|\right) \Big),$$

où le NDCG évalue l'accord dans l'importance classée des caractéristiques entre deux détecteurs, et $|\cdot|$ correspond à la valeur absolue. Si les deux modèles attribuent une importance SHAP élevée aux mêmes caractéristiques, leur valeur NDCG sera proche de 1, indiquant un comportement explicatif similaire [2].

3.3 Lier les similarités des explications aux sorties des détecteurs

Pour comparer la similarité des explications des modèles avec la similarité des sorties, nous introduisons deux matrices supplémentaires : la matrice des corrélations de scores et la matrice des similarités de Jaccard.

Nous définissons la similarité entre deux modèles M_i et M_j sur la base de la corrélation de leurs scores d'anomalie. Plus précisément, nous calculons la corrélation de Pearson moyenne par instance entre leurs vecteurs de scores comme suit :

$$\rho_{ij}^{\text{Score}} = \frac{1}{n}\sum_{k=1}^{n} \text{corr}\big(s_i^{(k)}, s_j^{(k)}\big).$$

Il en résulte une matrice symétrique qui reflète la similarité par paires entre les modèles vis-à-vis des scores attribués.
Pour comparer directement les prédictions de deux modèles i et j, la similarité de Jaccard entre les deux vecteurs de prédictions a_i et a_j est calculée comme suit :

$$J_{ij} = \frac{|a_i \cap a_j|}{|a_i \cup a_j|},$$

avec $|\cdot|$ le cardinal de l'ensemble. Cette métrique mesure le chevauchement entre les modèles : une valeur de 1 implique des prédictions identiques, tandis que 0 implique des ensembles disjoints d'anomalies détectées.
Chaque matrice de similarité $P \in \{\rho^{PS}, \rho^{NDCG}, \rho^{Scores}, J\}$ peut également être transformée en une matrice de dissimilarité $D \in \{\delta^{PS}, \delta^{NDCG}, \delta^{Scores}, \delta^{J}\}$. Ces matrices sont calculées selon la relation $D = 1 - P$, où chaque élément représente la distance entre deux détecteurs.
Afin de quantifier la relation entre les différentes mesures, nous utilisons le test de Mantel [15] pour déterminer si deux matrices de dissimilarité données sont statistiquement corrélées. Le coefficient de Mantel (r_M) est calculé comme la corrélation de Pearson entre les éléments triangulaires supérieurs des matrices, la signification statistique étant établie par des tests de permutation.

4 Expérimentations

Nous avons mené des expériences pour répondre aux questions suivantes : (1) La similarité des explications implique-t-elle une similarité des prédictions ? (2) Les métriques basées sur SHAP surpassent-elles les sorties brutes pour quantifier la diversité ? et (3) Dans quelle mesure cette diversité impacte-t-elle la précision et la robustesse de l'ensemble résultant ?

4.1 Configuration expérimentale

Dans les expériences, 14 algorithmes UAD ont été utilisés : COF, KNN, LOF, IForest, PCA, CBLOF, LODA, HBOS, MCD, OCSVM, DAGMM, DeepSVDD, COPOD et ECOD tels qu'implémentés dans la bibliothèque PyOD [23]. Les hyperparamètres définis par les auteurs de chaque modèle ont été conservés et aucune indication du pourcentage d'anomalie n'a été fournie pour aucun jeu de données. Au niveau de l'explicabilité, nous utilisons l'implémentation de la méthode agnostique Kernel SHAP disponible dans la bibliothèque Python `shap`. Le choix de cette méthode était nécessaire pour assurer un cadre unifié d'explication à travers nos divers algorithmes. De plus, pour assurer une approximation stable, le jeu de données d'arrière-plan a été résumé en utilisant l'algorithme de partitionnement k-means avec $k = 50$ centroïdes.

Pour augmenter la robustesse des résultats, chaque jeu de données a été divisé à cinq reprises en un ensemble d'entraînement et un ensemble de test, l'ensemble d'entraînement représentant 80 % des données. La graine aléatoire pour chaque division a été fixée égale à l'indice d'itération pour assurer la reproductibilité. Le code source utilisé dans les expériences est disponible sur GitHub [1].

4.2 Jeux de données considérés

En raison du coût de calcul élevé associé à SHAP, la sélection des jeux de données a été restreinte aux 50 % plus petits jeux de données disponibles dans ADBench [9]. De plus, les jeux de données contenant plus de 20 caractéristiques ont été supprimés pour maintenir la faisabilité des calculs. Ce processus de filtrage initial a abouti à un ensemble de 16 jeux de données provenant de diverses applications : annthyroid (AN), breastw (BR), glass (GL), Hepatitis (HE), Lymphography (LY), mammography (MA), PageBlocks (PB), Pima (PI), Stamps (ST), thyroid (TH), vertebral (VE), vowels (VO), WBC (WB), Wilt (WL), wine (WN) et yeast (YE).

4.3 Corrélation entre les matrices de similarité

Nous avons calculé les quatre matrices de similarité ρ^{PS} (corrélation linéaire entre les SHAP), ρ^{NDCG} (similarité des classements d'importance des caractéristiques SHAP), ρ^{Scores} (corrélation linéaire entre les scores d'anomalie), J (indice de Jaccard entre les prédictions d'anomalie) pour chaque jeu de données. Les matrices moyennes sur tous les jeux de données sont présentées dans la Figure 1, où un partitionnement hiérarchique a été appliqué pour optimiser l'ordre. Visuellement, deux groupes de modèles émergent. Premièrement, COPOD et ECOD se regroupent systématiquement, ce qui est attendu car les deux algorithmes reposent sur des hypothèses de distribution de données similaires. Deuxièmement, un groupe plus large comprenant OCSVM, AutoEncoder, IForest, PCA, KNN, CBLOF et GMM présente une forte corrélation. Ces algorithmes partagent des caractéristiques sous-jacentes liées aux métriques de distance et aux stratégies d'encodage des données.
Le test de Mantel est utilisé pour évaluer si les matrices de similarité sont statistiquement corrélées. Le Tableau 1 rapporte les corrélations de Mantel moyennes sur les jeux de données. Nous observons une forte corrélation ($r_M = 0.83$) entre δ^{PS} et δ^{NDCG}, confirmant que la similarité basée sur SHAP est cohérente en termes de magnitude et de classement de l'importance des caractéristiques. De même, δ^{Scores} et δ^{J} montrent une forte corrélation ($r_M = 0.76$), ce qui est intuitif puisque les modèles avec des distributions de scores similaires ont tendance à produire des prédictions binaires similaires. Enfin, la corrélation entre δ^{PS} et δ^{J} ($r_M = 0.67$) suggère que les détecteurs partageant des schémas de raisonnement similaires (tels que capturés par SHAP) ont également tendance à donner des prédictions d'anomalie similaires. Cependant, la similarité entre les modèles dépend du jeu de données. Plus précisément, le Tableau 2 montre les corrélations

1. https://github.com/jordanlv/Analyzing-SHAP-UOMS

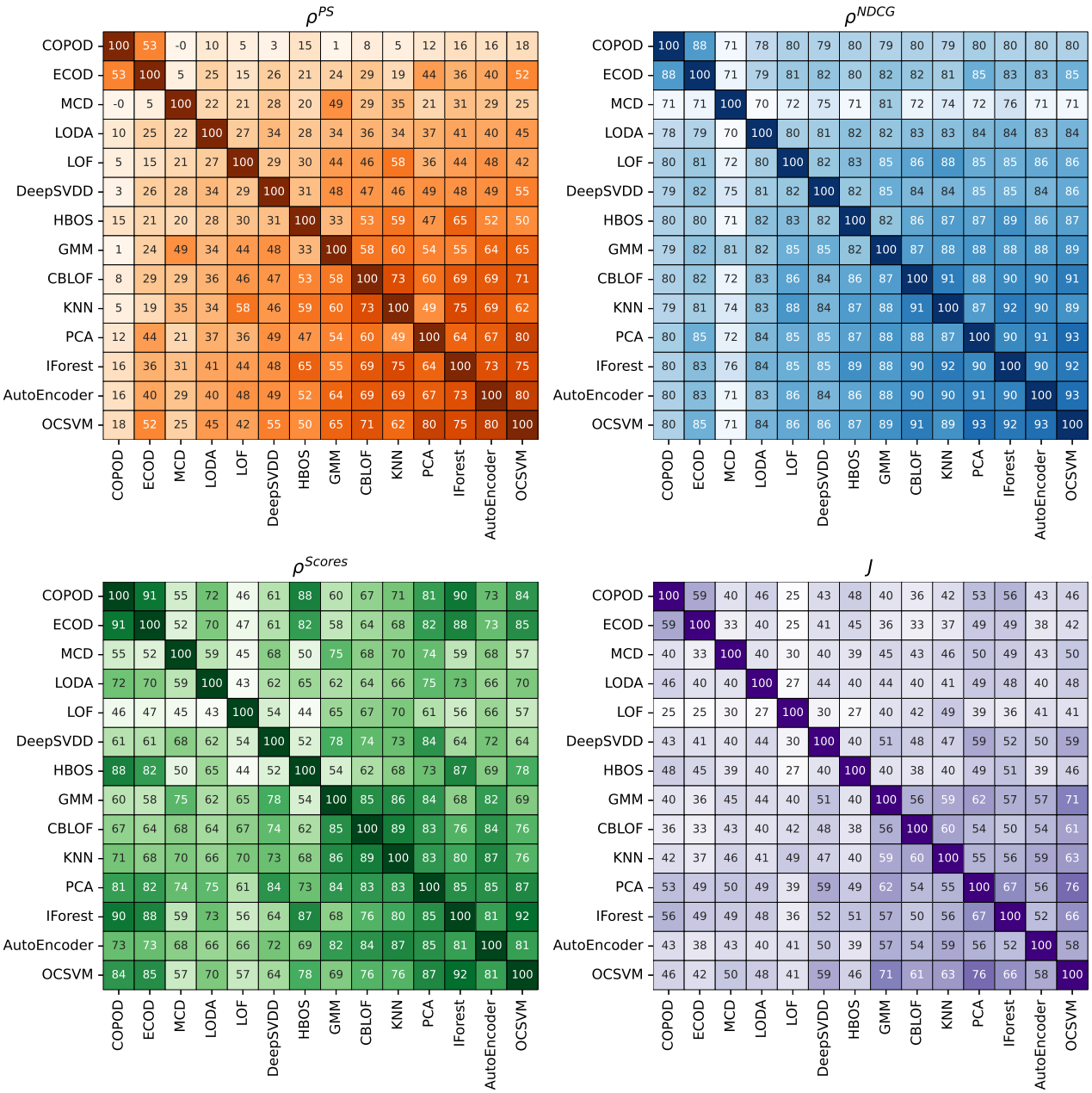


FIGURE 1 – Similarité moyenne entre les modèles sur tous les jeux de données. Disposition : Haut-Gauche : corrélations des valeurs SHAP ; Haut-Droite : NDCG des SHAP ; Bas-Gauche : corrélations des scores d'anomalie ; et Bas-Droite : similarités de Jaccard.

entre δ^{PS} et δ^{Scores}, et entre δ^{PS} et δ^{J} pour chaque jeu de données. Habituellement, les deux paires sont corrélées. Certains jeux de données montrent de fortes corrélations entre les matrices ($r_M > 0.5$) comme AN, PI, TH ou VO, tandis que d'autres ont une corrélation modérée ou faible entre les deux matrices. Enfin, certains jeux de données intéressants sont BR, LY ou YE où $r_M(\delta^{PS}, \delta^{Scores}) \ll r_M(\delta^{PS}, \delta^{J})$.

Dans l'ensemble, la similarité des modèles est corrélée avec leurs prédictions. Dans les sections suivantes, nous étudions la combinaison de modèles éloignés dans les matrices de similarité pour améliorer les résultats globaux d'un ensemble.

4.4 Agrégation des prédictions des modèles

L'objectif d'une méthode ensembliste est de combiner les prédictions de plusieurs détecteurs pour créer une méthode plus robuste. Un défi dans ce type de méthode est de savoir comment agréger chaque prédiction. Chaque détecteur d'anomalies produit un score d'anomalie et une prédiction indiquant si un point de données est normal ou anormal. Pour tirer véritablement parti de la spécialité de chaque modèle, nous utilisons les scores d'anomalie comme entrée

TABLE 1 – Corrélations de Mantel moyennes entre les matrices de distance, moyennées sur les jeux de données. Toutes les corrélations sont significatives ($p \leq 0.004$).

	δ^{PS}	δ^{NDCG}	δ^{Scores}	δ^{J}
δ^{PS}	1.00	0.83	0.55	0.67
δ^{NDCG}		1.00	0.54	0.57
δ^{Scores}			1.00	0.76
δ^{J}				1.00

TABLE 2 – Corrélations de Mantel ($\times 10^2$) entre les matrices de distance pour chaque jeu de données.

Jeu de données	$r_M(\delta^{PS}, \delta^{Scores})$	$r_M(\delta^{PS}, \delta^{J})$
AN	75	78
BR	22	51
GL	42	24
HE	26	51
LY	36	84
MA	56	51
PB	33	66
PI	55	52
ST	56	53
TH	72	76
VE	26	-4
VO	56	58
WB	18	15
WL	49	49
WN	26	30
YE	38	62

pour notre fonction d'agrégation. Les méthodes d'agrégation connues pour ces scores incluent l'utilisation du maximum, de la moyenne arithmétique, ou la moyenne des rangs (une méthode consistant à substituer les scores bruts par leur classement d'anormalité avant de les moyenner) [1]. Nous avons évalué ces fonctions sur tous les ensembles possibles de 3 modèles distincts parmi notre groupe de 14 modèles (résultant en 364 ensembles) sur chaque jeu de données. Compte tenu du déséquilibre de classe de chaque jeu de données, nous utilisons l'Aire Sous la Courbe Précision-Rappel (AUCPR) pour évaluer les performances. Comme le montre le Tableau 3, l'agrégation par rang donne des résultats supérieurs sur 11 des 16 jeux de données. Par conséquent, nous adoptons l'agrégation par rang pour le reste de cette étude.

4.5 Complémentarité

Pour sélectionner les modèles les plus dissimilaires, nous utilisons les matrices de dissimilarité D, ce qui donne des matrices sur la distance entre deux détecteurs. De nouveau, à partir du groupe de 14 modèles, nous avons construit des ensembles de taille $n = 3$. Le Tableau 4 présente les corrélations entre l'AUCPR des ensembles et les distances de diversité pour chaque matrice de dissimilarité. Une corrélation positive entre la diversité et la performance implique que l'augmentation de la diversité de l'ensemble conduit à des résultats de détection supérieurs.

TABLE 3 – AUCPR moyen ($\times 10^2$) entre tous les 364 ensembles avec les 3 principales stratégies d'agrégation. Les meilleurs résultats sont en gras.

Jeu de données	Rang	Max	Moyenne
AN	**26**	21	9
BR	**98**	59	49
GL	18	**22**	19
HE	**84**	64	47
LY	**100**	68	36
MA	**30**	8	4
PB	**60**	35	12
PI	**55**	45	42
ST	**62**	35	22
TH	**58**	28	5
VE	18	**28**	**28**
VO	**41**	11	8
WB	**97**	41	25
WL	6	**9**	7
WN	49	**53**	31
YE	37	**40**	39
moyenne	**52**	35	24

TABLE 4 – Corrélation ($\times 10^2$) entre la distance des modèles et l'AUCPR des ensembles. Les plus grands résultats sont en gras.

Jeu de données	δ^{PS}	δ^{NDCG}	δ^{Scores}	δ^{J}
AN	-20	**39**	-1	-32
BR	**-0**	-17	-90	-59
GL	**5**	-2	-33	-37
HE	**-28**	-32	-48	-31
LY	15	-0	-29	**16**
MA	**-9**	-68	-18	3
PB	-18	**21**	-17	-4
PI	**37**	14	30	24
ST	**40**	15	15	36
TH	-38	**19**	-12	-48
VE	-44	-30	-12	**-7**
VO	-66	-54	**-0**	-3
WB	**51**	36	-11	-28
WL	55	**62**	46	55
WN	43	**48**	7	1
YE	29	37	34	**45**
moyenne	3	**5**	-9	-4

Le tableau présente trois informations clés. Premièrement, la sélection de la diversité à partir des valeurs SHAP tend

à donner de meilleures performances que les sorties des modèles (scores et prédictions). Sur 11 des 16 jeux de données, l'utilisation de δ^{PS} et δ^{NDCG} comme diversité tend à donner de meilleurs résultats que l'utilisation de δ^{Scores} et δ^{J}. Deuxièmement, comme indiqué précédemment, la diversité contribue à l'apprentissage d'ensemble en élargissant la gamme des anomalies détectées. Cependant, une sélection de modèles efficace doit également tenir compte de la performance individuelle des modèles, un facteur non explicitement optimisé dans cette étude. Cette limitation explique probablement pourquoi les corrélations entre les métriques de distance et la performance restent modérées. De plus, dans certains jeux de données, ces corrélations sont systématiquement négatives, suggérant que la diversité n'est pas toujours bénéfique. Ce phénomène se produit notamment lorsqu'un seul modèle est plus performant que les autres. Dans de tels cas, la condition principale pour un ensemble réussi est l'inclusion de ce modèle spécifique. De plus, des corrélations négatives peuvent survenir dans des jeux de données très complexes où tous les modèles présentent de mauvaises performances. Dans ces scénarios, même une grande diversité ne peut pas compenser le manque de détection significative, entraînant un ensemble inefficace. Enfin, il est intéressant de noter que les SHAP et les scores ne montrent pas la même diversité car les corrélations varient entre les jeux de données. Par exemple, sur le jeu de données WB, les corrélations entre δ^{Scores} et δ^{J} sont négatives, tandis que δ^{PS} et δ^{NDCG} sont positives. Par conséquent, les deux métriques mettent en évidence des diversités différentes.

4.6 Diversité et performances individuelles

Bien que nous ayons montré que la diversité améliore la performance de l'ensemble en élargissant la couverture des anomalies, notre analyse précédente a négligé la précision individuelle des modèles. Ici, nous affinons nos résultats pour démontrer que malgré la valeur de la diversité, la qualité individuelle de chaque modèle reste un facteur critique. Pour les besoins de cette section, nous utilisons δ^{PS} pour quantifier la diversité.

La Figure 2 illustre l'importance de la performance individuelle de chaque modèle pour obtenir un ensemble efficace. Plus précisément, la Figure 2a présente les résultats pour le jeu de données LY. Une corrélation claire entre la diversité, la performance individuelle et la précision de l'ensemble est observable sur ce jeu de données. Cependant, cette corrélation n'est pas toujours bénéfique. Dans certains scénarios, imposer la diversité peut être préjudiciable à l'ensemble. Par exemple, la Figure 2b révèle que pour le jeu de données VO, la diversité offre un gain négligeable et peut même conduire à des résultats sous-optimaux.

Pour évaluer l'impact relatif de la qualité du modèle par rapport à la diversité, nous avons effectué une régression linéaire pour prédire le gain de performance de l'ensemble en utilisant la performance individuelle moyenne et les scores de diversité. Les poids résultants, présentés dans le Tableau 5, indiquent que bien que la performance individuelle soit généralement le facteur dominant, la diversité joue un rôle complémentaire crucial. Dans 12 des 16 jeux de données, le coefficient de diversité est positif, confirmant sa valeur en tant qu'amplificateur de performance. Notamment, pour les jeux de données WB et LY, le poids de la diversité rivalise ou dépasse même celui de la performance individuelle (ratios de 1,2 et 0,8 respectivement). Dans l'ensemble, avec un ratio moyen de 0,2, la diversité joue un rôle secondaire mais précieux dans la conception d'ensembles UAD.

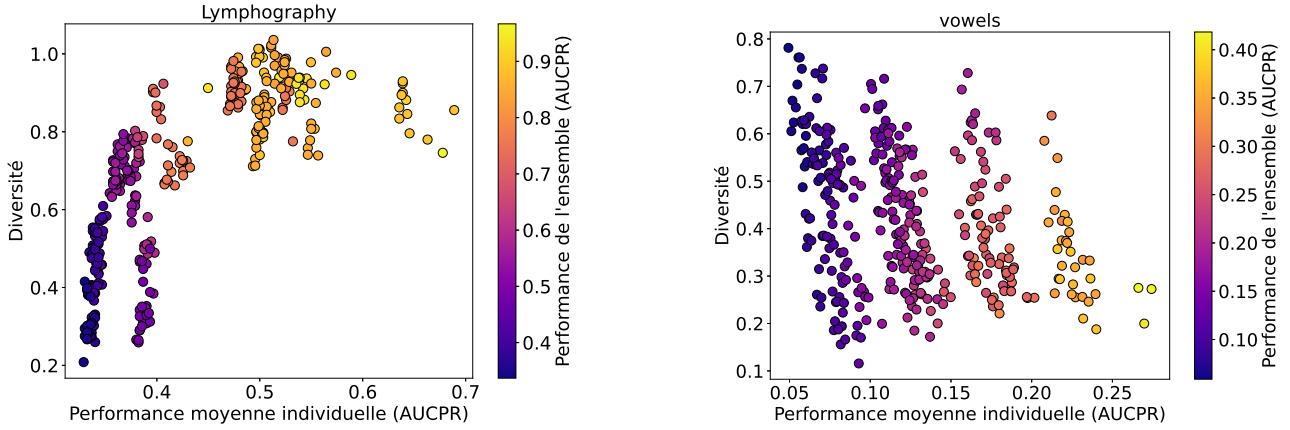
TABLE 5 – Poids de régression linéaire ($\times 10^2$) prédisant la performance de l'ensemble à partir de la performance individuelle moyenne et de la diversité, ainsi que leur ratio (vert : positif ; rouge : négatif).

Jeu de données	Perf. Indiv.	Diversité	Ratio
AN	2.7	0.5	0.2
BR	6.3	2.0	0.3
GL	1.1	0.1	0.1
HE	6.5	0.5	0.1
LY	10.7	8.4	0.8
MA	3.4	0.7	0.2
PB	5.0	2.2	0.4
PI	1.7	0.2	0.1
ST	5.6	0.3	0.0
TH	5.7	1.7	0.3
VE	0.1	-0.0	-0.1
VO	7.5	-0.6	-0.1
WB	2.4	2.8	1.2
WL	0.0	0.0	0.4
WN	8.0	-0.1	-0.0
YE	0.3	-0.0	-0.0
moyenne	4.2	1.2	0.2

5 Conclusion

Nous avons présenté une méthodologie pour sélectionner des modèles dans des ensembles UAD, basée sur la similarité de leurs explications. Nous avons démontré que la diversité s'avère bénéfique et permet d'améliorer les résultats. Nous avons analysé quatre métriques de diversité : deux fondées sur les explications SHAP et deux directement sur les sorties des modèles. Ces métriques ont conduit à des résultats différents lors de la sélection, indiquant qu'elles capturent des formes de diversité distinctes. De manière générale, les métriques basées sur SHAP ont affiché des résultats supérieurs à ceux basés sur les sorties. Enfin, nous avons établi que malgré les avantages de la diversité, la performance individuelle des modèles reste le facteur décisif : des modèles faibles mais diversifiés ne peuvent pas surpasser des modèles forts mais similaires. Cette recherche tend à montrer que l'explicabilité devrait être davantage prise en compte dans l'UOMS, car elle fournit de nouvelles informations sur le comportement des modèles.

Une limite de notre approche réside dans le coût de calcul de



(a) Jeu de données Lymphography (LY).

(b) Jeu de données Vowels (VO).

FIGURE 2 – Relation entre la diversité de l'ensemble (donnée par δ^{PS}) et la performance individuelle moyenne. Chaque point représente un ensemble de modèles. L'échelle de couleurs indique la performance globale de l'ensemble (AUCPR).

SHAP, qui peut devenir prohibitif pour les jeux de données comportant un grand nombre d'instances ou de caractéristiques. Cependant, le coût de calcul peut être atténué en utilisant des techniques d'approximation ou des modèles de substitution. De plus, notre stratégie est agnostique quant à la méthode d'explication, permettant l'utilisation de techniques d'interprétabilité moins coûteuses si nécessaire. Par exemple, dans [4], les auteurs ont récemment démontré comment l'agrégation de profils de dépendance partielle (PDP) sur un ensemble de modèles quasi-optimaux peut fournir des métriques fiables pour l'incertitude et la robustesse des explications.

Bien que nous ayons modélisé avec succès la diversité, l'optimisation des modèles individuels constitue une perspective d'évolution naturelle de ces travaux. L'ajustement fin des hyperparamètres permettrait de maximiser le potentiel de chaque détecteur, offrant ainsi un levier supplémentaire pour améliorer la performance globale de l'ensemble.

Concernant les travaux futurs, nous envisageons d'étudier la divergence entre les similarités basées sur SHAP et celles issues des sorties brutes afin d'affiner le processus de sélection de modèles. Par ailleurs, la performance individuelle étant critique, l'intégration de méthodes d'estimation de la qualité des modèles au sein de notre méthodologie constitue une priorité. Enfin, nous visons à étendre cette stratégie à l'UAD pour les séries temporelles, un domaine où la complexité accrue rend les approches ensemblistes particulièrement pertinentes [10].

Remerciements

La recherche présentée dans cet article a bénéficié d'un financement de l'Association Nationale de la Recherche et de la Technologie sous le numéro de subvention CIFRE 2023/1398 et de la société Soben. Les auteurs remercient également l'Agence Nationale de la Recherche pour le financement du projet MIMICO sous le numéro de subvention ANR-24-CE23-0380.

Références

[1] Charu C Aggarwal. *Outlier Analysis*. Springer, 2016.

[2] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[3] David Campos, Tung Kieu, Chenjuan Guo, Feiteng Huang, Kai Zheng, Bin Yang, and Christian S Jensen. Unsupervised time series outlier detection with diversity-driven convolutional ensembles. *Proceedings of the VLDB Endowment*, 15(3) :611–623, 2021.

[4] Mustafa Cavus, Jan N van Rijn, and Przemysław Biecek. Beyond the single-best model : Rashomon partial dependence profile for trustworthy explanations in automl. In *International Conference on Discovery Science*, pages 445–459. Springer, 2025.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection : A survey. *ACM computing surveys (CSUR)*, 41(3) :1–58, 2009.

[6] Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. *arXiv preprint arXiv :1905.12614*, 2019.

[7] Moncef Garouani, Ayah Barhrhouj, and Olivier Teste. Xstacking : An effective and inherently explainable framework for stacked ensemble learning. *Information Fusion*, 2025.

[8] Nicolas Goix. How to evaluate the quality of unsupervised anomaly detection algorithms ? *arXiv preprint arXiv :1607.01152*, 2016.

[9] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. ADBench : anomaly detection benchmark. *NeurIPS*, 35 :32142–32159, 2022.

[10] Jordan Levy, Clément Blanco-Volle, Nicolas Verstaevel, Benoit Gaudou, and Vincent Talon. Timeciel : Contextual interactive ensemble learning for time series classification. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 316–327. Springer, 2025.

[11] Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. InfoGAN-CR and ModelCentrality : Self-supervised Model Training and Selection for Disentangling GANs. In *International conference on machine learning*, pages 6127–6139. PMLR, 2020.

[12] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *NeurIPS*, 30, 2017.

[13] Martin Q Ma, Yue Zhao, Xiaorong Zhang, and Leman Akoglu. The need for unsupervised outlier model selection : A review and evaluation of internal evaluation strategies. *ACM SIGKDD Explorations Newsletter*, 25(1) :19–35, 2023.

[14] Andreas Madsen, Himabindu Lakkaraju, Siva Reddy, and Sarath Chandar. Interpretability needs a new paradigm. *arXiv*, 2024.

[15] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27 :209–220, 1967.

[16] Henrique O Marques, Ricardo JGB Campello, Jörg Sander, and Arthur Zimek. Internal evaluation of unsupervised outlier detection. *TKDD*, 14(4) :1–42, 2020.

[17] Sang Won Oh, Hye Seon Jo, Ho Jun Lee, Man Gyun Na, SW Oh, HS Jo, HJ Lee, and MG Na. Anomalies detection by unsupervised learning using explainable artificial intelligence in nuclear power plants. In *Transactions of the Korean Nuclear Society Spring Meeting Jeju, Korea*, 2022.

[18] Shebuti Rayana and Leman Akoglu. Less is more : Building selective anomaly ensembles. *TKDD*, 10(4) :1–33, 2016.

[19] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5) :756–795, 2021.

[20] Paul Saves, Pramudita Satria Palar, Muhammad Daffa Robani, Nicolas Verstaevel, Moncef Garouani, Julien Aligon, Benoit Gaudou, Koji Shimoyama, and Joseph Morlier. Surrogate modeling and explainable artificial intelligence for complex systems : A workflow for automated simulation exploration. *arXiv preprint*, 2025.

[21] Erich Schubert, Remigius Wojdanowski, Arthur Zimek, and Hans-Peter Kriegel. On evaluation of outlier rankings and outlier scores. In *International conference on data mining*, pages 1047–1058. SIAM, 2012.

[22] David H Wolpert and William G Macready. No free lunch theorems for optimization. *Transactions on evolutionary computation*, 2002.

[23] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod : A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96) :1–7, 2019.

Apprentissage Multi-Échelle Global-Local de Représentations de Nœuds par Marches Aléatoires et Optimisation Guidée par Points de Repère

Ali Assi[1], Nour Elislem Karabadji[2], Mohamed Elati[3], Wajdi Dhifli[3]

[1] Rochester Institute of Technology, Dubai, United Arab Emirates
[2] National Higher School of Technology and Engineering, Annaba, Algeria
[3] Univ. Lille, Inserm, CHU Lille, CNRS, U1366-UMR9020 – CRCLille – Cancer Research Center of Lille, Lille, F-59000, France

ali.assi@rit.edu, n.karabadji@ensti-annaba.dz,
mohamed.elati@univ-lille.fr, wajdi.dhifli@univ-lille.fr

Résumé

L'apprentissage de représentations vectorielles (embeddings) de nœuds en faible dimension est fondamental pour diverses tâches d'apprentissage sur graphes. Nous introduisons GLOW*, une approche optimisant conjointement : 1) l'information structurelle locale et contextuelle, capturée via des marches aléatoires, et 2) le positionnement global, approximé par des nœuds de repère (landmarks). La méthode utilise la divergence de Kullback-Leibler pour aligner ces deux aspects dans l'espace des représentations. Évaluée sur la classification de nœuds et la prédiction de liens,* GLOW *démontre des améliorations substantielles par rapport à l'état de l'art. Cet article est un résumé de notre publication à CIKM 2025 [2].*

Mots-clés

Apprentissage de représentations, Embedding de nœuds, Marches aléatoires, Classification, Prédiction de liens.

Abstract

Learning low-dimensional node representations is fundamental to graph learning tasks. We introduce GLOW*, an approach that jointly optimizes : 1) local structural information captured through random walks, and 2) global positioning approximated using landmark nodes. Evaluated on node classification and link prediction,* GLOW *demonstrates substantial improvements over the state-of-the-art. This paper is a summary of our publication at CIKM 2025 [2].*

Keywords

Representation learning, Node embedding, Random walk, Link prediction, Node classification.

1 Introduction

Les graphes sont largement adoptés pour modéliser des données relationnelles dans divers domaines (réseaux sociaux, molécules, réseaux de citations, etc.). L'apprentissage de représentations (*embeddings*) de nœuds en faible dimension est crucial pour des tâches telles que la classification de nœuds et la prédiction de liens. De nombreuses méthodes récentes s'appuient sur des marches aléatoires [1, 3, 7] pour capturer des voisinages locaux, ou sur les réseaux de neurones sur graphes (GNN) [4, 8] pour propager l'information entre nœuds voisins. Toutefois, ces approches peinent souvent à préserver simultanément les proximités locales et le positionnement global des nœuds dans le graphe.

Pour surmonter ces limites, nous proposons GLOW (**G**lobal-**L**ocal **O**ptimization for random **W**alk and landmark-based node embeddings) [2][1], un algorithme non supervisé qui intègre l'information structurelle locale capturée via un modèle de langage appliqué sur des marches aléatoires [6], et affine ces représentations en utilisant un ensemble stratégiquement sélectionné de nœuds de repère (landmarks) pour approximer le positionnement global.

2 Apprentissage de représentations avec GLOW

Soit un graphe $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. L'objectif de GLOW est d'apprendre une fonction d'*embedding* $f : \mathcal{V} \to \mathbb{R}^d$ ($d \ll |\mathcal{V}|$) qui préserve la structure locale et globale du graphe.

2.1 Information locale par marches aléatoires

Un ensemble de marches aléatoires est généré pour chaque nœud $v \in \mathcal{V}$. Ces séquences sont traitées comme des phrases dans un modèle sémantique de type Skip-gram [6], produisant une matrice de représentations initiale $\mathbf{E} \in \mathbb{R}^{n \times d}$ capturant le voisinage direct (1-hop) et contextuel (N-hop) de chaque nœud.

1. https://github.com/wajdidhifli/GLOW

2.2 Positionnement global par nœuds de repère

Un sous-ensemble $\mathcal{L} \subset \mathcal{V}$ de nœuds de repère ($|\mathcal{L}| \ll |\mathcal{V}|$) est sélectionné de manière diversifiée : le premier landmark est le nœud de plus haut degré dans chaque composante connexe, puis les suivants sont choisis en maximisant la distance aux landmarks déjà sélectionnés. Pour les graphes à composantes déconnectées, un nœud "pont" est ajouté, relié au landmark de plus haut degré de chaque composante, afin de permettre le calcul des plus courts chemins inter-composantes (avec un terme de pénalité γ pour préserver la séparation structurelle).

Les distances des plus courts chemins entre chaque nœud et les landmarks forment une matrice $\mathbf{SG} \in \mathbb{R}^{n \times |\mathcal{L}|}$, représentant le positionnement global. Une matrice correspondante $\mathbf{SEG} \in \mathbb{R}^{n \times |\mathcal{L}|}$ est calculée à partir des distances cosinus entre les *embeddings* des nœuds et ceux des landmarks dans l'espace $\mathbf{E}$. Après normalisation L1 ligne par ligne des deux matrices, la divergence de Kullback-Leibler (KL) [5] est utilisée comme fonction de perte :

$$Loss_{KL}(\mathbf{E}) = \sum_{i=1}^{|\mathcal{V}|} \sum_{j=1}^{|\mathcal{L}|} \mathbf{SG}_{ij} \log \left(\frac{\mathbf{SG}_{ij}}{\mathbf{SEG}_{ij}} \right) \qquad (1)$$

En minimisant cette perte par descente de gradient, les représentations se raffinent itérativement pour aligner les relations de similarité dans l'espace d'*embedding* avec les distances structurelles du graphe.

3 Évaluation expérimentale

GLOW est évalué sur la classification de nœuds (Accuracy) et la prédiction de liens (AUC), sur sept jeux de données réels.

Prédiction de liens. Le Tableau 1 présente les résultats AUC (moyenne sur 10 exécutions) sur Cora, Citeseer et Pubmed. GLOW obtient systématiquement les meilleurs scores, surpassant toutes les approches concurrentes, y compris les méthodes GNN récentes (SEAL, NBF-net, HPLC) et les approches par marches aléatoires (NESS).

TABLE 1 – AUC de prédiction de liens (% ± écart-type).

Approche	Cora	Citeseer	Pubmed
DeepWalk	83.10 ± 0.50	80.50 ± 0.80	84.40 ± 0.92
Node2Vec	84.49 ± 0.49	80.00 ± 0.68	80.32 ± 0.29
GCN (GAE)	88.68 ± 0.40	85.35 ± 0.60	95.80 ± 0.13
GraphSAGE	90.24 ± 0.34	87.37 ± 1.39	96.58 ± 0.11
SEAL	92.55 ± 0.50	85.82 ± 0.44	95.86 ± 0.28
HPLC	94.95 ± 0.18	96.15 ± 0.19	97.38 ± 0.34
NESS	98.13 ± 0.20	99.43 ± 0.10	96.60 ± 0.10
GLOW	**98.64** ± 0.22	**99.76** ± 0.17	**99.64** ± 0.02

Classification de nœuds. GLOW obtient les meilleures précisions sur 6 des 7 jeux de données évalués, surpassant des modèles GNN de pointe (GCN, GAT). Par exemple, GLOW atteint 85.8% sur Cora, 80.2% sur Citeseer, 93.3% sur Coauthor Physics et 88.4% sur DBLP (soit $+12.1\%$ par rapport à la deuxième meilleure méthode SGC sur ce dernier). La seule exception est Coauthor CS où GCN obtient 91.0% contre 90.4% pour GLOW.

4 Conclusion

Nous avons présenté GLOW, une méthode d'*embedding* de nœuds qui fusionne les informations locales de voisinage et le positionnement global via une optimisation basée sur la divergence KL et des nœuds de repère. Nos résultats démontrent que GLOW surpasse de nombreuses méthodes de l'état de l'art en classification de nœuds et en prédiction de liens, confirmant l'intérêt de modéliser conjointement les dépendances structurelles multi-échelle.

Références

[1] Ali Assi and Wajdi Dhifli. Instance matching in knowledge graphs through random walks and semantics. *Future Generation Computer Systems*, 123 :73–84, 2021.

[2] Ali Assi, Nour Elislem Karabadji, Mohamed Elati, and Wajdi Dhifli. Learning global-local multi-scale node embeddings with random walks and landmark-guided optimization. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*, pages 79–89, 2025.

[3] Aditya Grover and Jure Leskovec. node2vec : Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, 2016.

[4] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[5] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22 :79–86, 1951.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 3111–3119, 2013.

[7] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk : Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710, 2014.

[8] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Explications Probabilistes Préférées Pareto-Optimales : Équilibrer Contraintes Cognitives et Préférences Utilisateur

Louenas Bounia[1]

[1] LIPN-UMR CNRS 7030 Université Sorbonne Paris Nord, Villetaneuse, France

bounia@lipn.univ-paris13.fr

Résumé

*Les modéles d'apprentissage automatique déployés dans des domaines critiques nécessitent des explications compréhensibles. Les explications abductives identifient des ensembles minimaux de caractéristiques garantissant une décision, mais souffrent de deux limitations : dépasser les limites cognitives humaines et ignorer les préférences utilisateur (actionnabilité, équité, coût). Les explications probabilistes réduisent la taille via une erreur contrôlée, tandis que les explications préférées intègrent les préférences ; cependant, ces approches restent disjointes. Nous introduisons le premier cadre unifiant ces paradigmes : des explications probabilistes préférées équilibrant contraintes cognitives, précision probabiliste et préférences utilisateur. Nous formulons le problème via une scalarisation pondérée (*WPPE*) et une optimisation de Pareto (*PPPE*), prouvons la* NP*-dureté pour les arbres de décision, et exploitons la supermodularité pour établir des garanties d'approximation. Nous proposons trois algorithmes complémentaires : la descente gloutonne pondérée (*WGD*) avec ratio $(e^{p_w} - 1)/p_w$, l'énumération de la frontière de Pareto (*PFE*) pour l'exploration interactive des compromis non-dominés, et l'algorithme lexicographique stratifié (*LSA*) pour les préférences ordinales strictes. Notre cadre permet une navigation rigoureuse dans l'espace des compromis Pareto-optimaux entre limites cognitives et préférences utilisateur.*

Mots-clés

Explications abductives, Explications probabilistes, Optimisation multi-objectifs, Pareto-optimalité, Préférences utilisateur, IA explicable.

Abstract

*Machine learning classifiers deployed in critical domains require comprehensible explanations. Abductive explanations identify minimal feature sets guaranteeing a decision but suffer from two limitations : exceeding human cognitive limits and ignoring user preferences (actionability, fairness, cost). Probabilistic explanations reduce size via controlled error, while preferred explanations integrate preferences ; however, these approaches remain disjoint. We introduce the first framework unifying these paradigms : preferred probabilistic explanations that balance cognitive constraints, probabilistic accuracy, and user preferences. We formulate the problem via weighted scalarization (*WPPE*) and Pareto optimization (*PPPE*), prove* NP*-hardness for decision trees, and exploit supermodularity to establish approximation guarantees. We propose three complementary algorithms : weighted greedy descent (*WGD*) with ratio $(e^{p_w} - 1)/p_w$, Pareto frontier enumeration (*PFE*) for interactive exploration of non-dominated trade-offs, and lexicographic stratified algorithm (*LSA*) for strict ordinal preferences. Our framework enables rigorous navigation of the Pareto-optimal trade-off space between cognitive limits and user preferences.*

Keywords

Abductive explanations, Probabilistic explanations, Multi-objective optimization, Pareto optimality, User preferences, Explainable AI.

1 Introduction

1.1 Motivation et Contexte

Les Classifieurs d'apprentissage automatique sont de plus en plus déployés dans des domaines à enjeux élevés tels que la santé [21], la justice pénale [19], l'octroi de prêts et les décisions d'emploi [47]. Ces modèles, incluant les arbres de décision, les forêts aléatoires et les réseaux de neurones, ont démontré des performances prédictives remarquables dans diverses applications.

Cependant, avec leur déploiement dans des secteurs critiques, la capacité d'expliquer les décisions des modèles est devenue essentielle pour la transparence, la confiance et la conformité réglementaire (droit à l'explication du RGPD, loi européenne sur l'IA). Ce besoin a conduit au développement de l'IA Explicable (XAI), avec l'adoption généralisée de méthodes agnostiques post-hoc populaires telles que LIME [49], SHAP [37] et les explications contrefactuelles [57]. Ces méthodes reposent sur des perturbations locales ou la théorie des jeux sans tenir compte de la structure du modèle.

Malgré leur popularité, les méthodes agnostiques souffrent de limitations fondamentales : elles peuvent générer des explications identiques pour des prédictions opposées [52, 23, 22], manquent de fondements théoriques rigoureux [35], et présentent une sensibilité aux perturbations des entrées [1].

Ces faiblesses compromettent la fiabilité dans les contextes critiques où les décisions ont des conséquences significatives.

Notre travail s'inscrit dans le cadre de l'XAI Formelle [38], fournissant des explications mathématiquement validées avec des garanties de fidélité et de robustesse. La forme la plus prominente est celle des *explications abductives*, qui identifient des ensembles minimaux de caractéristiques dont la fixation des valeurs garantit la même prédiction indépendamment des autres caractéristiques. Chaque caractéristique est nécessaire et collectivement suffisante pour justifier la décision.

Cependant, deux défis interdépendants émergent :

Défi 1 : Surcharge Cognitive. Même les explications minimales peuvent contenir trop de caractéristiques pour être traitées cognitivement. [41] ont établi que les humains ne peuvent traiter simultanément que 7 ± 2 éléments, une limite affinée à seulement 4 éléments pour les tâches de raisonnement complexe [14].

Défi 2 : Hétérogénéité des Préférences. Toutes les caractéristiques ne sont pas également pertinentes, et la même explication peut convenir à un utilisateur mais pas à un autre. L'utilité d'une explication dépend du profil utilisateur, de son expertise et du contexte de décision, notamment en ce qui concerne l'interprétabilité, l'actionnabilité et les contraintes éthiques [42]. Ignorer cette hétérogénéité peut conduire à des explications formellement correctes mais non informatives pour les utilisateurs finaux, comme l'illustrent les scénarios suivants :

— **Scénario A (Diagnostic médical) :** Un patient non expert préfère des explications avec des termes médicaux courants, tandis qu'un médecin est à l'aise avec cette terminologie.

— **Scénario B (Demande de prêt) :** Certaines caractéristiques, telles que *âge* $>$ 50, ne sont pas actionnables, tandis que d'autres, comme le *ratio dette/revenu* $>$ 0,6, peuvent être modifiées.

— **Scénario C (Applications sensibles au coût) :** Dans les contextes industriels et médicaux, certaines variables reposent sur des capteurs ou des tests coûteux, limitant leur utilisation dans les explications.

1.2 Le Manque dans la Recherche Actuelle

Les travaux existants traitent la concision et la compréhensibilité de manière disjointe. Les **explications probabilistes** [26, 58, 31, 34, 59, 12, 13, 2] relaxent l'exactitude via une erreur contrôlée ϵ, réduisant la taille pour respecter les limites cognitives, mais traitent toutes les caractéristiques uniformément sans intégrer les préférences utilisateur. À l'inverse, les **explications préférées** [4] intègrent les préférences (coût, actionnabilité, équité) [42, 53, 36] mais requièrent $\epsilon = 0$, produisant de longues explications mal adaptées aux besoins diversifiés des parties prenantes [36].

1.3 Contributions

Nous proposons le premier cadre raisonné unifiant les explications probabilistes et les préférences utilisateur. Notre approche calcule des explications qui sont simultanément **courtes** (via une erreur contrôlée ϵ), **précises** (avec des garanties probabilistes explicites) et **alignées sur les préférences** (optimisant l'utilité).

Contributions algorithmiques. Nous développons trois algorithmes complémentaires exploitant la supermodularité : la **Descente Gloutonne Pondérée (WGD)** optimise un objectif scalarisé avec un ratio d'approximation $(e^{p_w} - 1)/p_w$; l'**Énumération de la Frontière de Pareto (PFE)** explore efficacement les compromis non-dominés sans recherche exhaustive ; et l'**Algorithme Lexicographique Stratifié (LSA)** gère les préférences ordinales via des niveaux de priorité avec des garanties en temps polynomial.

Contributions théoriques. Nous établissons : (i) la NP-difficile de WPPE pour les arbres de décision, (ii) la préservation de la supermodularité sous pondération des préférences — cruciale pour nos garanties d'approximation, et (iii) la tractabilité en temps polynomial pour les arbres de décision malgré la dureté générale, permettant une implémentation pratique.

2 Préliminaires

Problèmes de classification. Nous considérons une classification binaire booléenne avec d caractéristiques. Soit $[d] = \{1, \ldots, d\}$ les indices des caractéristiques et $X_d = \{x_1, \ldots, x_d\}$ l'ensemble des caractéristiques. Chaque instance $\boldsymbol{x} = (x_1, \ldots, x_d) \in \{0,1\}^d$ est classifiée par $h : \{0,1\}^d \to \{0,1\}$. Nous supposons que le lecteur est familier avec l'apprentissage supervisé et les arbres de décision.

Explications Abductives et Probabilistes. Une *explication abductive*[1] pour $\boldsymbol{x}$ est un sous-ensemble de caractéristiques $S \subseteq [d]$ tel que toute instance $\boldsymbol{y}$ coïncidant avec $\boldsymbol{x}$ sur S satisfait $h(\boldsymbol{y}) = h(\boldsymbol{x})$. Une *raison suffisante* est une explication abductive minimale. Ces explications pouvant être longues, les *explications probabilistes* [2, 59] relaxent la correction via une erreur contrôlée. La *fonction d'erreur* $\mu_{h,x}(S)$ compte les instances $\boldsymbol{y}$ où $y_S = x_S$ mais $h(\boldsymbol{y}) \neq h(\boldsymbol{x})$. L'*erreur normalisée*

$$\epsilon_{h,x}(S) = \frac{\mu_{h,x}(S)}{2^{d-|S|}}$$

mesure la proportion de complétions incompatibles. Un ensemble S est une *explication probabiliste de niveau* ϵ si $\epsilon_{h,x}(S) \leq \epsilon$.

Explications Préférées. Toutes les explications ne sont pas également utiles. Les *explications préférées* [4] tiennent compte des préférences utilisateur (actionnabilité, coût, contraintes) [42, 56] via des pondérations de caractéristiques ou des préférences ordinales. Nous utilisons des

1. Contrairement à [23], nous ne nous restreignons pas aux explications abductives minimales au sens de l'inclusion ensembliste.

coûts additifs : x_i reçoit un poids $w(x_i) \in R^+$, donnant $\text{cost}(S) = \sum_{i \in S} w(x_i)$.

Fonctions Sous-modulaires. Une fonction $f : 2^{[d]} \to R$ est *monotone* si $f(S) \leq f(T)$ (croissante) ou $f(S) \geq f(T)$ (décroissante) pour tout $S \subseteq T$. Elle est *supermodulaire* si la perte marginale $L_f(i \mid S) = f(S) - f(S \setminus \{i\})$ satisfait $L_f(i \mid S) \geq L_f(i \mid T)$ pour tout $S \subseteq T$ et $i \in S$, intuitivement, retirer un élément impacte davantage les petits ensembles. Une fonction est *sous-modulaire* si l'inverse est vérifié ; *modulaire* si les contributions sont additives. La *courbure* mesure l'écart à la modularité et est cruciale pour les garanties d'approximation. Pour f positive et $I \subseteq [d]$ non-vide : $c_f(I) = 1 - \min_{i \in I} \frac{f(I) - f(I \setminus \{i\})}{f(\{i\}) - f(\emptyset)}$. On a $c_f(I) \in [0, 1]$, où $c_f(I) = 0$ indique la modularité et $c_f(I) = 1$ la super/sous-modularité maximale. Pour f supermodulaire décroissante, $c_f(I) < 1$ garantit une approximation à facteur constant sous contraintes de cardinalité [24, 55].

2.1 Travaux Connexes

Notre travail s'inscrit dans un courant de recherche connu sous le nom d'*IA explicable formelle* (XAI formelle) [38, 17, 16], qui étudie les explications dotées de garanties prouvables [51, 5, 8, 11, 10].

Explications Abductives et Probabilistes. [16] ont caractérisé les raisons suffisantes pour diverses classes de modèles, avec des algorithmes efficaces pour les arbres de décision [27, 2, 13], les forêts aléatoires [25, 30], les modèles linéaires [39, 54], les ensembles d'arbres [7] et les réseaux de neurones [11, 60]. Sur le versant probabiliste, Waldchen [61] ont prouvé la $\mathsf{NP}^{\mathsf{PP}}$-complétude pour les circuits booléens, Arenas et al [2] ont montré la NP-complétude pour les arbres de décision, et [12] ont exploité la sous-modularité pour l'approximation avec des garanties formelles.

Explications Préférées et Multi-Objectifs. [42, 48] souligne l'importance de l'interprétabilité personnalisée. [4] ont intégré les contraintes de coût, d'actionnabilité et d'équité [56, 32], mais se restreignent aux explications exactes ($\epsilon = 0$). Notre travail généralise cette approche en autorisant $\epsilon > 0$ via un compromis paramétré par λ, géré par WGD pour les préférences quantitatives et LSA pour les préférences ordinales. [44] a considéré l'optimisation multi-objectifs pour les contrefactuels ; notre PFE étend cela aux explications probabilistes sur l'espace de Pareto (ϵ, cost), en s'appuyant sur l'optimisation sous-modulaire [45, 55] pour une approximation efficace des problèmes NP-difficiles.

3 Formulations du Problème

3.1 Problème Principal

Le nombre d'explications abductives peut croître exponentiellement avec les caractéristiques, même pour des modèles simples comme les arbres de décision et les ensembles d'arbres [43, 27, 3, 29], rendant l'analyse exhaustive impraticable. Certaines sont longues et difficiles à interpréter [28, 6], tandis que d'autres peuvent ne pas refléter les préférences utilisateur. Pour résoudre ce compromis, nous considérons les *explications probabilistes préférées*, qui autorisent une erreur probabiliste contrôlée ϵ pour réduire la taille tout en intégrant explicitement les préférences utilisateur sur les caractéristiques.

Définition 3.1 (Explications Probabilistes Préférées)**.** *Étant donné un Classifieur $h : \{0,1\}^d \to \{0,1\}$, une instance $x \in \{0,1\}^d$ et une fonction de poids $w : X_d \to R^+$ modélisant les préférences utilisateur, une explication probabiliste préférée est un sous-ensemble $S \subseteq [d]$ minimisant l'erreur probabiliste $\epsilon_{h,x}(S)$ tout en optimisant le coût $\text{cost}(S) = \sum_{i \in S} w(x_i)$ sous la contrainte $|S| \leq k$.*

Le problème central calcule des sous-ensembles de taille au plus k réalisant des compromis optimaux entre précision et satisfaction des préférences, conduisant aux formulations par scalarisation pondérée et optimisation de Pareto ci-dessous.

3.2 Problème de Scalarisation Pondérée (WPPE)

La première approche agrège les deux objectifs — erreur probabiliste et coût de préférence — en une seule fonction objectif via un paramètre de compromis λ.

Problème 3.2 (**WPPE** – Explications Probabilistes Préférées Pondérées)**.** ***Entrée :***

- *Un Classifieur $h : \{0,1\}^d \to \{0,1\}$*
- *Une instance $x \in \{0,1\}^d$ à expliquer*
- *Un ensemble candidat de caractéristiques $I \subseteq [d]$*
- *Une limite de taille $k \leq |I|$ reflétant les contraintes cognitives*
- *Une fonction de poids $w : X_d \to R^+$ encodant les préférences*
- *Un paramètre de compromis $\lambda \geq 0$ contrôlant l'importance relative du coût*

Sortie : *Un sous-ensemble de caractéristiques S^* résolvant*

$$S^* \in \underset{S \subseteq I, |S| \leq k}{\text{argmin}} \; [\epsilon_{h,x}(S) + \lambda \cdot \text{cost}(S)]$$

Interprétation : Le paramètre λ agit comme un taux de change entre les objectifs : $\lambda = 0$ privilégie exclusivement la minimisation de l'erreur (en ignorant les préférences), tandis qu'un grand λ favorise les explications à faible coût au détriment de la précision. Puisque nous minimisons le coût de préférence, les caractéristiques non préférées reçoivent des poids élevés tandis que les caractéristiques préférées reçoivent des poids faibles. Cette formulation est particulièrement adaptée lorsque les utilisateurs peuvent exprimer explicitement leurs préférences sous forme de compromis quantitatif, comme c'est souvent le cas dans les contextes médicaux.

Remarque 3.3 (Cas particuliers et complexité)**.** WPPE *généralise les explications probabilistes classiques : lorsque $\lambda = 0$, ou lorsque les préférences sont uniformes ($w(x_i) = c$ pour tout i) et que la taille $|S|$ est fixée exactement à*

k, le problème se réduit à $\min_{S\subseteq I,|S|\leq k} \epsilon_{h,x}(S)$. Puisque ce problème est $\mathsf{NP}^{\mathsf{PP}}$*-complet pour les classifieurs* CNF *[61]—strictement au-dessus de* NP—WPPE *hérite de cette difficulté.*

Proposition 3.4 (Complexité)**.** *Le problème* WPPE *est* NP*-difficile lorsque le Classifieur h est représenté par un arbre de décision et I est une raison suffisante.*

3.3 Problème d'Optimisation de Pareto (PPPE)

La seconde formulation évite de choisir λ a priori en considérant l'ensemble complet des compromis optimaux, appelé frontière de Pareto [20]. Contrairement à la scalarisation pondérée qui agrège les deux objectifs en une fonction scalaire, l'approche de Pareto minimise simultanément l'erreur probabiliste $\epsilon_{h,x}(\cdot)$ et le coût de préférence $\mathrm{cost}(\cdot)$ sans privilégier l'un ou l'autre.

Définition 3.5 (Dominance de Pareto)**.** *Soient deux explications $S, S' \subseteq I$ avec $|S|, |S'| \leq k$. On dit que S* domine au sens de Pareto *S', noté $S \prec_P S'$, si et seulement si :*

$$\epsilon_{h,x}(S) \leq \epsilon_{h,x}(S') \quad \textit{et} \quad \mathrm{cost}(S) \leq \mathrm{cost}(S'),$$

avec au moins une inégalité stricte.

Définition 3.6 (Optimalité de Pareto)**.** *Une explication $S \subseteq I$ avec $|S| \leq k$ est* Pareto-optimale *si aucune autre explication $S' \subseteq I$ de taille au plus k ne la domine, c'est-à-dire qu'il n'existe pas de S' tel que $S' \prec_P S$. L'ensemble de toutes les explications Pareto-optimales constitue la* frontière de Pareto*, notée $\mathcal{P}_k$.*

Problème 3.7 (PPPE – Explications Probabilistes Préférées de Pareto)**.** ***Entrée :***

- *Un Classifieur $h : \{0,1\}^d \to \{0,1\}$*
- *Une instance $x \in \{0,1\}^d$ à expliquer*
- *Un ensemble candidat de caractéristiques $I \subseteq [d]$*
- *Une limite de taille $k \leq |I|$ reflétant les contraintes cognitives*
- *Une fonction de poids $w : X_d \to R^+$ encodant les préférences*

Sortie : *La frontière de Pareto $\mathcal{P}_k$, ou une approximation de celle-ci, i.e., l'ensemble (ou sur-ensemble) de toutes les explications Pareto-optimales de taille au plus k par rapport aux objectifs de minimisation $\epsilon_{h,x}(\cdot)$ et $\mathrm{cost}(\cdot)$.*

Avantages de l'approche de Pareto. Cette formulation offre plusieurs avantages. Premièrement, elle ne nécessite aucune spécification a priori du compromis, adaptée aux préférences mal définies ou évolutives [40]. Deuxièmement, elle permet une exploration interactive : les utilisateurs examinent la frontière $\mathcal{P}_k$ et sélectionnent leur compromis préféré. Troisièmement, elle révèle les *points de coude* [18] où les taux de substitution erreur-coût changent brusquement. Enfin, contrairement à la scalarisation pondérée qui capture un seul point par λ, elle fournit une caractérisation complète des solutions non-dominées.

4 Algorithmes d'Approximation

Nous présentons maintenant trois algorithmes complémentaires pour résoudre efficacement les problèmes WPPE et PPPE, en exploitant systématiquement la structure mathématique sous-jacente.

4.1 Propriétés Structurelles

Notre approche algorithmique repose fondamentalement sur une propriété de supermodularité relaxée de $\mu_{h,x}$ établie dans [12], en notant que $\epsilon_{h,x}$ n'est ni supermodulaire ni monotone en général.

Lemme 4.1 (Préservation de la Supermodularité)**.** *Soit $\mu_{h,x} : 2^{[d]} \to R^+$ une fonction supermodulaire et monotone décroissante. Soit $w : X_d \to R^+$ une fonction de poids. Définissons la fonction objectif pondérée :*

$$f_\lambda(S) = \mu_{h,x}(S) + \lambda \sum_{i\in S} w(x_i)$$

Alors $f_\lambda(\cdot)$ est supermodulaire, positive et monotone décroissante pour tout $\lambda \geq 0$ satisfaisant :

$$\lambda \leq \lambda^* = \min_{S\subseteq[d],\, i\notin S} \frac{\mu_{h,x}(S) - \mu_{h,x}(S\cup\{i\})}{w(x_i)}$$

4.2 Algorithme 1 (WGD)

Notre premier algorithme résout WPPE via une stratégie de descente gloutonne, en retirant itérativement les caractéristiques qui minimisent l'augmentation de la fonction objectif pondérée.

Algorithm 1 Descente Gloutonne Pondérée (WGD)

Require: Classifieur h, instance x, ensemble initial I, limite k, poids w, paramètre λ
Ensure: Explication $S_{\text{WGD}} \subseteq I$ avec $|S_{\text{WGD}}| \leq k$

```
n ← |I|, S_n ← I
for j = n jusqu'à 1 do
    for chaque caractéristique i ∈ S_j do
        Calculer la marge : Δ_i = f_λ(S_j \ {i}) − f_λ(S_j)
    end for
    i* ← argmin_{i∈S_j} Δ_i
    S_{j−1} ← S_j \ {i*}
    Stocker la paire (S_j, ε_{h,x}(S_j)) pour évaluation ultérieure
end for
retourner S_WGD ∈ argmin_{S_j : |S_j| ≤ k} [ε_{h,x}(S_j) + λ · cost(S_j)]
```

Intuition algorithmique. L'algorithme démarre avec l'ensemble complet I (qui peut être une raison suffisante) et retire itérativement la caractéristique dont la suppression cause la plus petite dégradation de la fonction objectif pondérée. Ce processus génère une séquence emboîtée $I = S_n \supset S_{n-1} \supset \cdots \supset S_1 \supset S_0 = \emptyset$, et l'algorithme retourne finalement la meilleure solution parmi celles-ci.

Proposition 4.2 (Garantie d'Approximation). *Soit $\lambda \leq \lambda^*$ tel que défini dans le Lemme 4.1, de sorte que f_λ est supermodulaire, monotone décroissante et positive sur 2^I. Soit $S^* \in \operatorname{argmin}_{S \subseteq I, |S| \leq k} f_\lambda(S)$ une solution optimale de* WPPE*, et soit c_w la courbure de f_λ sur 2^I. Alors l'algorithme* WGD *retourne une solution S_{WGD} satisfaisant :*

$$f_\lambda(S_{WGD}) = \epsilon_{h,x}(S_{WGD}) + \lambda \cdot \text{cost}(S_{WGD}) \leq \alpha_w \cdot f_\lambda(S^*)$$

où $\alpha_w = \frac{e^{p_w}-1}{p_w}$ avec $p_w = \frac{c_w}{1-c_w}$.

Remarque 4.3. *En pratique, les ratios d'approximation observés sont significativement meilleurs que les bornes théoriques, comme le montrent nos expériences (Section 5). Cela s'explique par le fait que les bornes théoriques sont établies pour le pire cas, tandis que les instances réelles présentent généralement une structure plus favorable.*

4.3 Algorithme 2 : Énumération de la Frontière de Pareto (PFE)

Pour résoudre **PPPE** et construire une approximation de la frontière de Pareto, nous proposons une stratégie d'échantillonnage systématique de l'espace des paramètres λ.

Algorithm 2 Énumération de la Frontière de Pareto (PFE)

Require: Classifieur h, instance x, ensemble I, limite k, poids w, nombre d'échantillons L
Ensure: Frontière de Pareto approchée $\tilde{\mathcal{P}}$

Définir $\lambda_{\min} = 10^{-3}$
Calculer $\lambda^* = \min_{S \subseteq [d],\, i \notin S} \frac{\mu_{h,x}(S) - \mu_{h,x}(S \cup \{i\})}{w(x_i)}$
Calculer le facteur géométrique : $\gamma = \left(\frac{\lambda^*}{\lambda_{\min}}\right)^{1/(L-1)}$
Générer la séquence $\{\lambda_1, \ldots, \lambda_L\}$ avec $\lambda_1 = \lambda_{\min}$ et $\lambda_{\ell+1} = \gamma \cdot \lambda_\ell$
Solutions $\leftarrow \emptyset$
for chaque $\lambda \in \{\lambda_1, \ldots, \lambda_L\}$ **do**
    $S_\lambda \leftarrow \text{WGD}(h, x, I, k, w, \lambda)$
    Calculer $\epsilon_\lambda = \epsilon_{h,x}(S_\lambda)$ et $c_\lambda = \text{cost}(S_\lambda)$
    Solutions $\leftarrow$ Solutions $\cup \{(S_\lambda, \epsilon_\lambda, c_\lambda)\}$
end for
Supprimer les solutions dupliquées (paire (ϵ, c))
$\tilde{\mathcal{P}} \leftarrow$ FilterDominance(Solutions)
return $\tilde{\mathcal{P}}$

Procédure FilterDominance. Cette procédure itère sur l'ensemble des solutions et élimine celles dominées au sens de Pareto. Pour chaque paire de solutions (S, S'), elle vérifie si S domine S' (auquel cas S' est supprimée) ou vice versa. Elle peut être implémentée efficacement en temps $O(L^2)$ en triant préalablement les solutions selon une dimension.

Choix de la progression géométrique. L'utilisation d'une suite géométrique pour λ (plutôt qu'arithmétique) est motivée par deux considérations. Premièrement, elle assure une exploration uniforme de l'espace des compromis sur échelle logarithmique sur $[\lambda_{\min}, \lambda^*]$, ce qui est naturel car les ratios de coût sont souvent plus significatifs que leurs différences absolues. Deuxièmement, elle garantit une couverture complète de la frontière de Pareto dans le domaine de validité du Lemme 4.1, comme formalisé par le théorème suivant.

Définition 4.4 (Approximation de Pareto (α, β)). *Une frontière de Pareto approchée $\tilde{\mathcal{P}}$ est une approximation (α, β) de la vraie frontière de Pareto $\mathcal{P}$ si, pour toute solution Pareto-optimale $S^* \in \mathcal{P}$, il existe une solution $\tilde{S} \in \tilde{\mathcal{P}}$ satisfaisant :*

$$\epsilon_{h,x}(\tilde{S}) \leq \alpha \cdot \epsilon_{h,x}(S^*) \quad \textit{et} \quad \text{cost}(\tilde{S}) \leq \beta \cdot \text{cost}(S^*)$$

Théorème 4.5 (Garantie pour PFE). *Soit α_w le ratio d'approximation de* WGD *tel que défini dans la Proposition 4.2, sous la condition $\lambda \leq \lambda^*$ définie dans le Lemme 4.1. Alors l'algorithme* PFE *avec L échantillons tirés de $[\lambda_{\min}, \lambda^*]$ retourne une frontière $\tilde{\mathcal{P}}$ qui est une approximation de Pareto (α_w, α_w) de $\mathcal{P}$.*

4.4 Algorithme 3 : Approche Lexicographique Stratifiée (LSA)

Dans certains contextes applicatifs, les préférences utilisateur ne sont pas quantitatives mais ordinales : les utilisateurs peuvent classer les caractéristiques par niveaux de priorité, souhaitant utiliser en premier les caractéristiques de niveau 1, puis celles de niveau 2 uniquement si nécessaire, et ainsi de suite. Cette situation est fréquente en médecine (tests invasifs vs non-invasifs) ou en finance (données publiques vs privées).

Modèle de préférence lexicographique. Supposons que les caractéristiques soient partitionnées en m strates $I = I_1 \cup I_2 \cup \cdots \cup I_m$ avec $I_i \cap I_j = \emptyset$ pour $i \neq j$, où I_1 contient les caractéristiques les plus préférées et I_m les moins préférées. Une explication S est lexicographiquement préférée à S' si, en notant $n_i(S) = |S \cap I_i|$ le nombre de caractéristiques de S dans la strate i, il existe un niveau ℓ tel que :

$$n_i(S) = n_i(S') \quad \forall i < \ell \qquad \text{et} \qquad n_\ell(S) < n_\ell(S')$$

Principe de fonctionnement. L'algorithme LSA procède par niveaux de priorité, en commençant par les caractéristiques les moins préférées (niveau m) et en remontant progressivement vers les plus préférées (niveau 1). À chaque niveau ℓ, il retire itérativement des caractéristiques de I_ℓ tant que :

1. La limite de taille k n'est pas atteinte
2. L'erreur probabiliste reste inférieure à $\epsilon_{\max}$
3. Des caractéristiques restent à retirer dans le niveau courant

Ce processus garantit que les caractéristiques de faible priorité sont éliminées en premier, préservant autant que possible celles de haute priorité.

Théorème 4.6 (Garantie pour LSA). *Pour chaque niveau de priorité ℓ, l'algorithme LSA produit une solution qui est α_w-approchée parmi toutes les solutions utilisant au plus le même nombre de caractéristiques des niveaux 1 à $\ell - 1$.*

Algorithm 3 Algorithme Lexicographique Stratifié (LSA)

Require: Classifieur h, instance x, strates $\{I_1, \ldots, I_m\}$, limite k, seuil d'erreur $\epsilon_{\max}$
Ensure: Explication lexicographiquement optimale S_{LSA}

$S \leftarrow I_1 \cup I_2 \cup \cdots \cup I_m$
if $\epsilon_{h,x}(S) > \epsilon_{\max}$ **then**
  retourner S {Infaisable}
end if
for niveau $\ell = m$ **jusqu'à** 1 **do**
  while $S \cap I_\ell \neq \emptyset$ **do**
    $i^* \leftarrow \text{argmin}_{i \in S \cap I_\ell} [\mu_{h,x}(S \setminus \{i\}) - \mu_{h,x}(S)]$
    $S_{\text{temp}} \leftarrow S \setminus \{i^*\}$
    if $\epsilon_{h,x}(S_{\text{temp}}) \leq \epsilon_{\max}$ **et** $|S_{\text{temp}}| \geq 1$ **then**
      $S \leftarrow S_{\text{temp}}$
    else
      break
    end if
  end while
end for
retourner S

4.5 Application aux Arbres de Décision

Les garanties d'approximation de la Proposition 4.2 et des Théorèmes 4.5 et 4.6 s'appliquent à toute classe d'hypothèses booléenne. Cependant, pour l'efficacité computationnelle, chaque appel à l'oracle $\mu_{h,x}(S)$ doit s'exécuter en temps polynomial. L'évaluation de $\mu_{h,x}$ dans le cas général est #P-difficile [15]. Pour les arbres de décision, [27, 33, 9] ont montré qu'elle peut être calculée en temps polynomial, tandis que [12] a démontré une évaluation en temps linéaire $O(|T| \cdot |S|)$, où $|T|$ est le nombre de nœuds de l'arbre.

Complexités résultantes. Nos algorithmes héritent des complexités suivantes dans le pire cas pour les arbres de décision. WGD effectue $O(n^2)$ appels à l'oracle (où $n = |I|$), chacun nécessitant $O(|T| \cdot d)$ opérations, donnant $O(n^2 \cdot |T| \cdot d)$. PFE effectue L appels à WGD sur $[\lambda_{\min}, \lambda^*]$, donnant $O(L \cdot n^2 \cdot |T| \cdot d)$ avec typiquement $L \in [20, 50]$. LSA effectue au plus n itérations nécessitant chacune $O(n)$ évaluations de $\mu_{h,x}(\cdot)$, donnant la même complexité $O(n^2 \cdot |T| \cdot d)$ que WGD. Cette tractabilité est particulièrement pertinente car les arbres de décision sont largement utilisés dans les domaines critiques (santé, finance, justice) en raison de leur interprétabilité [50], et nos algorithmes étendent cela en permettant des explications plus concises alignées sur les préférences tout en préservant l'efficacité computationnelle.

4.6 Analyse Comparative

Ayant établi la tractabilité en temps polynomial pour les arbres de décision, nous comparons maintenant les trois algorithmes :

Recommandations d'utilisation :

— **WGD** : Préférer lorsque le compromis souhaité est bien défini ($\lambda \leq \lambda^*$) et qu'une solution unique est recherchée. Idéal pour l'intégration dans des systèmes automatisés.

— **PFE** : Recommandé pour l'exploration interactive, lorsque les préférences utilisateur ne sont pas entièrement spécifiées a priori, ou pour présenter plusieurs options à l'utilisateur final.

— **LSA** : À utiliser dans les contextes où les préférences sont naturellement hiérarchiques (médecine, finance réglementée) et où les compromis quantitatifs sont difficiles à exprimer.

TABLE 1 – Comparaison des trois algorithmes proposés

Critère	WGD	PFE	LSA
Type de préférence	Quantitatif	Quantitatif	Ordinal
Sortie	Une solution	Frontière complète	Une solution
Choix de λ requis	Oui ($\leq \lambda^*$)	Non	Non
Complexité	$O(n^2\|T\|d)$	$O(Ln^2\|T\|d)$	$O(n^2\|T\|d)$
Ratio d'approximation	α_w	(α_w, α_w)	α_w par niveau
Exploration interactive	Non	Oui	Non

5 ÉVALUATION EXPÉRIMENTALE

Pour valider l'efficacité et la scalabilité de nos trois algorithmes proposés (WGD, PFE, LSA), nous avons mené des expériences extensives sur des Classifieurs à arbres de décision sur divers jeux de données de référence. Tous les algorithmes ont été implémentés en Python, en exploitant scikit-learn pour l'entraînement des modèles et NumPy pour les calculs numériques. Les expériences ont été réalisées sur une station de travail équipée d'un processeur Intel Core i9-9900 à 3,1 GHz et 64 Gio de RAM, sous Ubuntu 22,04 LTS.

Jeux de données. Nous avons considéré $B = 25$ jeux de données de classification binaire (10 à 10^5 caractéristiques) issus de Kaggle, OpenML et UCI, incluant MNIST38 et MNIST49, deux sous-ensembles binaires de MNIST opposant les chiffres 3 vs. 8 et 4 vs. 9 respectivement. Les tâches multi-classes ont été converties en binaire en considérant la classe dominante contre toutes les autres. Par manque d'espace, nous reportons les résultats sur $B = 10$.

Préférences Utilisateur. Nous avons considéré quatre modèles de préférences : (i) **préférences uniformes**, où toutes les caractéristiques reçoivent un poids égal, servant de référence ; (ii) **préférences basées sur SHAP**, utilisant les valeurs de Shapley comme poids d'importance locaux ; (iii) **préférences basées sur LIME**, dérivées des explications locales de LIME, et (iv) **importance des caractéristiques**, utilisant l'importance des caractéristiques de scikit-learn pour les arbres de décision. Par manque d'espace, nous ne reportons que les résultats pour les préférences basées sur SHAP dans l'article principal. Les résultats pour les autres scénarios sont fournis en matériel supplémentaire.

5.1 PROTOCOLE EXPÉRIMENTAL

Tâches d'Explication. Pour chaque benchmark b, une tâche d'explication est $(T, \boldsymbol{x}, I, k, w)$. Le Classifieur h est

représenté par un arbre de décision T entraîné via l'implémentation CART de scikit-learn [46]. L'instance $\boldsymbol{x}$ est sélectionnée aléatoirement dans l'ensemble de test ; l'ensemble candidat I est le chemin racine-feuille cohérent avec $\boldsymbol{x}$ (explication de chemin [27]). La fonction de poids w est dérivée des scénarios de préférences via une transformation monotone décroissante : $w(x_i) = \max\{s\} - s_i + \delta$ où s_i est le score d'importance brut et $\delta > 0$ assure la positivité. Cela mappe une importance plus élevée à des poids plus faibles, ce qui est cohérent avec notre objectif de minimisation du coût. Nous fixons la limite cognitive $k = 7 \pm 2$ [41] en général ; lorsque $|I|$ est petit, nous adaptons et fixons $k = 4 \pm 2$ pour éviter les cas triviaux où $k \geq |I|$. Nous imposons une limite de temps de 60 minutes par instance. Les performances sont évaluées en moyennant $\epsilon_{h,x}(S)$, $\text{cost}(S)$ et $|S|$ sur $m = \min\{s, 150\}$ instances de test.

Scalabilité et Qualité d'Approximation. En dehors du cas $\epsilon = 0$, où les explications abductives préférées exactes peuvent être calculées via MAXSAT [4], et du cas $\lambda = 0$ où les explications probabilistes exactes sont accessibles via un encodage SAT [2], aucune méthode exacte n'est connue pour le problème général des explications probabilistes préférées. Cela motive notre approche par approximation. Pour évaluer empiriquement sa qualité, nous comparons l'Algorithme 1 (WGD) contre deux références exactes. Premièrement, pour $\lambda = 0$, nous utilisons la méthode basée sur SAT de [2], étendue à la version décisionnelle du Problème 3.4 en ajoutant la clause $\bigvee\{x_i : (x_I)_i = 1\} \vee \bigvee\{\neg x_i : (x_I)_i = 0\}$, et résolue via une recherche binaire sur $(0, 1]$ avec une précision 10^{-3}, nécessitant au plus 10 appels à GLUCOSE 4 (PySAT, timeout 60 min). Deuxièmement, pour $\lambda > 0$, nous comparons contre une référence par énumération exhaustive, calculant $\text{argmin}_{S \subseteq I, |S| \leq k} f_\lambda(S)$ sur tous les sous-ensembles possibles, sur des jeux de données de petite dimension où cela est tractable, comme *placement*, *tic-tac-toe* et d'autres.

Impact de λ sur WGD. Nous évaluons l'Algorithme 1 (WGD) sur $\lambda \in [10^{-3} \cdot 2^{d-k},\ 10^3 \cdot 2^{d-k}]$, où d est le nombre de caractéristiques et k la limite de taille, en utilisant des préférences normalisées basées sur SHAP avec $\sum_{i \in I} w(x_i) = 1$ pour assurer un taux d'échange significatif entre $\epsilon_{h,x}(\cdot)$ et $\text{cost}(\cdot)$.

Exploration de la Frontière de Pareto via PFE. Une limitation clé de WGD est qu'il nécessite de spécifier λ a priori, ce qui peut être difficile en pratique lorsque les préférences utilisateur ne sont pas entièrement quantifiées. L'Algorithme 2 (PFE) répond à cela en calculant un ensemble de solutions Pareto-optimales couvrant l'espace complet des compromis entre $\epsilon_{h,x}(\cdot)$ et $\text{cost}(\cdot)$, sans nécessiter de choix explicite de λ. Nous évaluons PFE avec $L = 30$ échantillons tirés de $\lambda \in [10^{-3}, 10^3]$ avec un pas géométrique, et reportons les frontières de Pareto approchées sous forme de graphiques 2D (ϵ vs. cost), mettant en évidence les points de coude où le taux de substitution marginal entre les deux objectifs change brusquement.

Préférences Ordinales et LSA. Tandis que WGD et PFE gèrent les préférences quantitatives via λ, certains domaines expriment naturellement les préférences en termes ordinaux. L'Algorithme 3 (LSA) répond à cela en partitionnant les caractéristiques en strates de priorité. Nous évaluons LSA sur COMPAS [56], et le comparons contre WGD en termes de $\epsilon_{h,x}(S)$, composition des strates $n_\ell(S)$ et $|S|$.

5.2 Résultats Expérimentaux

Le Tableau 2 reporte les résultats sur 10 des 25 benchmarks pour des Classifieurs à arbres de décision avec $k = 7$ et $\lambda = 0$. Les colonnes *acc* et d donnent la précision et le nombre de caractéristiques binaires. Les lignes sont triées par taille moyenne d'explication de chemin $|I|$ [27]. Nous reportons l'erreur moyenne $\epsilon_{h,\boldsymbol{x}}(S)$ et la taille d'explication $|S|$ pour WGD et la méthode exacte basée sur SAT, ainsi que le temps d'exécution de SAT. Les jeux de données en bleu indiquent un timeout partiel ; en magenta un timeout complet. WGD termine toujours en moins d'1 seconde. Pour le régime général $\lambda > 0$, le Tableau 3 traite le jeu de données *horse colic* sur 15 valeurs représentatives de $\lambda \in [10^{-3} \cdot 2^{d-k},\ 10^3 \cdot 2^{d-k}]$, comparant WGD contre l'énumération exhaustive ; les résultats confirment que WGD atteint un $f_\lambda(S)$ normalisé systématiquement inférieur à la méthode exacte pour λ petit et modéré, avec une borne d'approximation pratique bien en dessous de la garantie théorique de la Proposition 4.2. Cependant, pour de grandes valeurs de λ (au-delà de λ^*), les garanties théoriques ne tiennent plus et le ratio d'approximation peut se dégrader significativement. Cette dégradation est amplifiée lorsque I est une explication de chemin plutôt qu'une raison suffisante, car f_λ peut perdre sa structure de monotonicité dans ce régime, expliquant l'explosion occasionnelle de la borne observée dans le Tableau 3. Globalement, WGD démontre une scalabilité remarquable, restant stable en millisecondes sur de grands jeux de données comme *gisette* et *farm ads* où SAT dépasse le timeout — tandis que l'écart $\epsilon_{h,\boldsymbol{x}}(S_{\text{WGD}}) - \epsilon_{h,\boldsymbol{x}}(S^*)$ reste négligeable pour $\lambda = 0$ et $|S_{\text{WGD}}|$ est en moyenne inférieure à $|S^*|$.

Impact de λ sur WGD. La Figure 1 illustre l'effet de λ sur le jeu de données placement : à mesure que λ augmente, $\text{cost}(S)$ et $|S|$ décroissent monotonicalement tandis que $\epsilon_{h,x}(S)$ augmente, confirmant que λ contrôle efficacement le compromis entre satisfaction des préférences et précision probabiliste. Pour $\lambda \leq \lambda^*$, le ratio d'approximation empirique reste bien en dessous de la borne théorique de la Proposition 4.2, confirmant la quasi-optimalité de WGD. Au-delà de λ^*, les garanties d'approximation ne tiennent plus et la qualité des solutions peut se dégrader, en particulier lorsque I n'est pas une raison suffisante. Cela souligne la nécessité d'explorer l'espace complet des compromis via PFE plutôt que de s'engager sur un λ unique a priori.

Exploration de la Frontière de Pareto via PFE. La Figure 2 reporte la frontière de Pareto approchée retournée par PFE sur COMPAS (préférences SHAP, $k = 7$, $|I| = 11$), comprenant 7 solutions non-dominées allant de haute fidélité ($\epsilon = 0{,}274$, $\text{cost} = 0{,}998$, $|S| = 7$) à un ali-

Benchmark				$\epsilon_{h,\boldsymbol{x}}(S)$		$\lvert S\rvert$		Temps (s)
nom	*acc*	d	$\lvert I\rvert$	WGD	SAT	WGD	SAT	SAT
student perf.	92.04	30	5.37	0.27 (±0.10)	0.27 (±0.10)	2.03	2.03	2.11
hungarian	63.71	13	6.69	0.12 (±0.11)	0.12 (±0.09)	3.57	3.57	1.79
horse colic	75.94	40	6.74	0.14 (±0.06)	0.14 (±0.06)	4.09	4.09	11.71
loan eligibility	74.08	68	8.49	0.18 (±0.12)	0.22 (±0.13)	5.73	6.81	43.96
wine	70.11	11	9.05	0.09 (±0.09)	0.10 (±0.11)	5.66	5.64	35.48
employee attr.	83.37	63	10.58	0.07 (±0.08)	0.21 (±0.10)	6.44	7.04	1008.42
compas	68.02	40	10.97	0.05 (±0.08)	0.06 (±0.09)	5.86	6.82	1091.88
mnist49	96.28	784	15.60	0.38 (±0.13)	–	6.92	–	–
gisette	94.37	5000	21.39	0.33 (±0.10)	–	6.92	–	–
farm ads	81.12	54877	23.18	0.14 (±0.16)	–	6.34	–	–

TABLE 2 – Résultats expérimentaux pour $\lambda = 0$ (Problème 3.2) sur 10 benchmarks avec h arbre de décision, $k \in \{5, 6, 7\}$. En bleu : timeout partiel ; en magenta : timeout complet.

λ	$\epsilon_{h,\boldsymbol{x}}(S)$		$\lvert S\rvert$		$f_\lambda(S)$ (norm.)	
	WGD	Exact	WGD	Exact	WGD	Exact
0.0010	0.1105 (±0.0689)	0.1105 (±0.0689)	4.7387 (±0.6536)	4.7387 (±0.6536)	0.0587	0.0587
0.2507	0.1105 (±0.0689)	0.2268 (±0.0789)	4.7387 (±0.6536)	3.0270 (±1.3046)	0.0587	0.1205
0.5005	0.1122 (±0.0684)	0.2401 (±0.0719)	4.7297 (±0.6837)	2.8919 (±1.3244)	0.0596	0.1275
0.7502	0.1122 (±0.0684)	0.2626 (±0.0904)	4.7297 (±0.6837)	2.7658 (±1.2445)	0.0596	0.1395
10^0	0.1291 (±0.0614)	0.2672 (±0.0882)	4.6396 (±0.9280)	2.7387 (±1.2499)	0.0686	0.1419
2.15×10^6	0.2409 (±0.1115)	0.3825 (±0.1694)	3.1171 (±1.4316)	1.7748 (±0.9743)	0.1280	0.2032
6.98×10^6	0.2490 (±0.1174)	0.3905 (±0.1730)	3.0180 (±1.4205)	1.7027 (±0.9260)	0.1323	0.2074
1.18×10^7	0.2566 (±0.1157)	0.3944 (±0.1743)	2.9009 (±1.4203)	1.6847 (±0.9201)	0.1363	0.2095
1.66×10^7	0.2626 (±0.1214)	0.3949 (±0.1740)	2.8198 (±1.4155)	1.6847 (±0.9201)	0.1395	0.2098
2.15×10^7	0.2632 (±0.1230)	0.3949 (±0.1740)	2.8108 (±1.4111)	1.6847 (±0.9201)	0.1398	0.2098
2.15×10^{11}	0.3388 (±0.1526)	0.4342 (±0.1916)	2.0991 (±1.2003)	1.2432 (±0.5061)	0.2766	0.2306
3.82×10^{11}	0.3549 (±0.1562)	0.4351 (±0.1911)	2.0270 (±1.1505)	1.2252 (±0.4966)	0.3320	0.2311
6.79×10^{11}	0.3549 (±0.1562)	0.4400 (±0.1934)	2.0270 (±1.1505)	1.1982 (±0.4615)	0.4436	0.2337
1.21×10^{12}	0.3555 (±0.1561)	0.4402 (±0.1933)	2.0180 (±1.1546)	1.1802 (±0.4492)	0.6425	0.2338
2.15×10^{12}	0.3639 (±0.1581)	0.4428 (±0.1927)	1.9730 (±1.1347)	1.1622 (±0.3923)	1.0000	0.2352

TABLE 3 – WGD vs. méthode exacte sur Horse Colic ($d = 36$, $k = 5$, acc. $= 77{,}48\%$, préférences SHAP, $|R_d| \approx 6$, $\lambda \in [10^{-3} \cdot 2^{d-k},\ 10^3 \cdot 2^{d-k}]$). $f_{\tilde{\lambda}}(S)$ (avec $\tilde{\lambda} = \lambda / 2^{d-k}$).

gnement maximal des préférences ($\epsilon = 0{,}517$, cost ≈ 0, $|S| = 1$). Un point de coude particulièrement attractif est la deuxième solution de Pareto ($\epsilon = 0{,}285$, cost $= 0{,}119$, $|S| = 7$), qui réalise une réduction de 88% du coût de préférence au prix d'une augmentation marginale de l'erreur de $0{,}011$, illustrant comment PFE permet aux utilisateurs d'identifier des solutions fortement alignées sur leurs préférences sans connaissance explicite de λ.

Préférences Ordinales et LSA. L'Algorithme 3 (LSA) est évalué sur COMPAS [56] avec $\epsilon_{\max} = 0{,}1$, $k = 7$ et trois strates : I_1 (actionnables : *Number_of_Priors*, *score_factor*, *Age_**), I_2 (partiels : *Misdemeanor*), et I_3 (attributs démographiques non-actionnables). LSA élimine lexicographiquement trois caractéristiques non-actionnables de I_3 tout en préservant $\epsilon_{h,x}(S) = 0$, donnant $n_1 = 4$, $n_2 = 1$, $n_3 = 2$ avec $|S| = 7$. Cela est significatif dans le contexte COMPAS : l'explication repose principalement sur des caractéristiques actionnables (*Number_of_Priors*, *score_factor*) tout en minimisant les attributs démographiques sensibles. En contraste, WGD ($\lambda = 0$) retourne $|S| = 5$ avec seulement $n_1 = 3$ et $n_2 = 0$, ignorant complètement la structure ordinale. LSA atteint un ratio non-actionnable strictement inférieur à erreur égale, confirmant qu'il produit des explications à la fois probabilistiquement valides et maximalement alignées sur les priorités utilisateur.

6 Conclusion et Perspectives

Nous avons introduit le premier cadre unifié pour les *explications probabilistes préférées*, reliant précision probabiliste, contraintes cognitives et préférences utilisateur via trois algorithmes complémentaires : WGD pour les préférences quantitatives avec ratio d'approximation formel $(e^{p_w} - 1)/p_w$, PFE pour l'exploration interactive de la frontière de Pareto sans engagement a priori sur λ, et LSA pour les préférences ordinales lexicographiques. Les trois exploitent la structure supermodulaire de $\mu_{h,x}$ pour assurer une tractabilité en temps polynomial et des certificats de qualité formels pour les arbres de décision. Les expériences sur 25 benchmarks confirment que WGD monte en charge sur des jeux de données allant jusqu'à $54{,}877$ caractéristiques en millisecondes là où les méthodes exactes basées sur SAT dépassent le timeout, que PFE révèle des points de coude significatifs permettant aux utilisateurs de sélectionner interactivement leur compromis précision-coût préféré, et que LSA produit systématiquement des explications avec moins de caractéristiques non-actionnables que WGD à erreur probabiliste égale, comme démontré sur COMPAS.

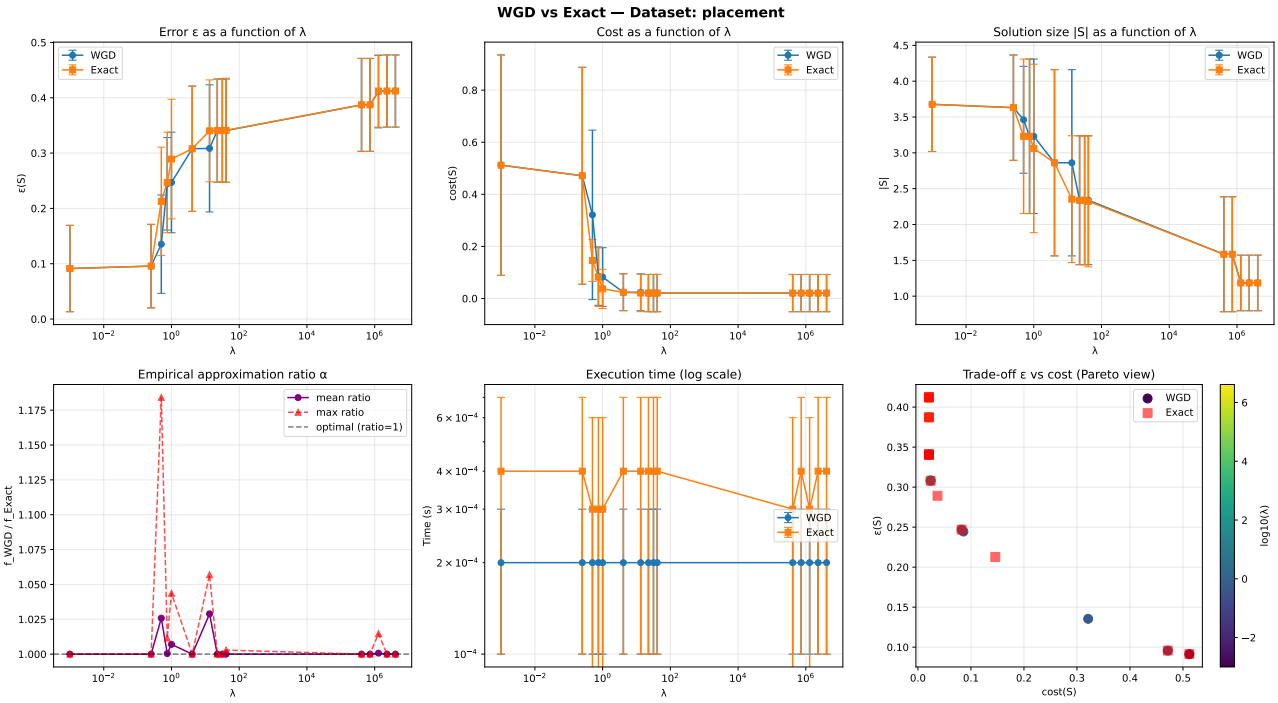


FIGURE 1 – WGD vs. méthode exacte sur le jeu de données placement (préférences SHAP) et $k = 4$.

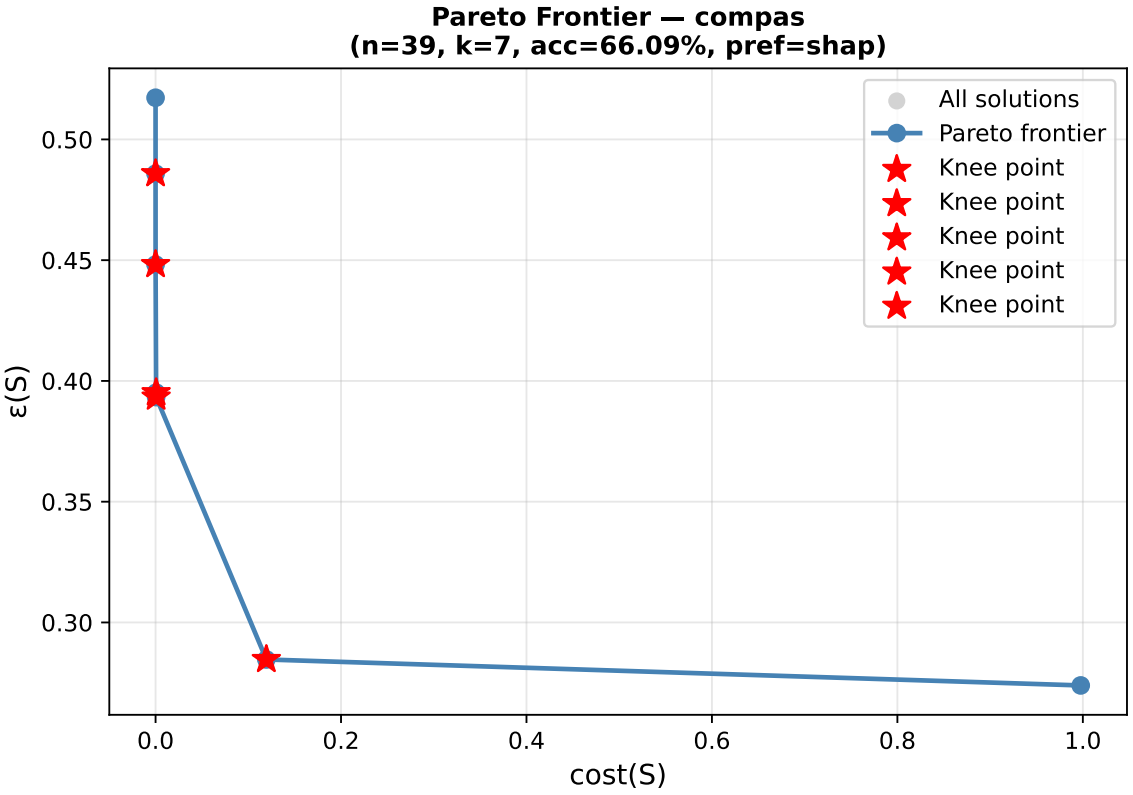


FIGURE 2 – Frontière de Pareto obtenue par PFE sur COMPAS.

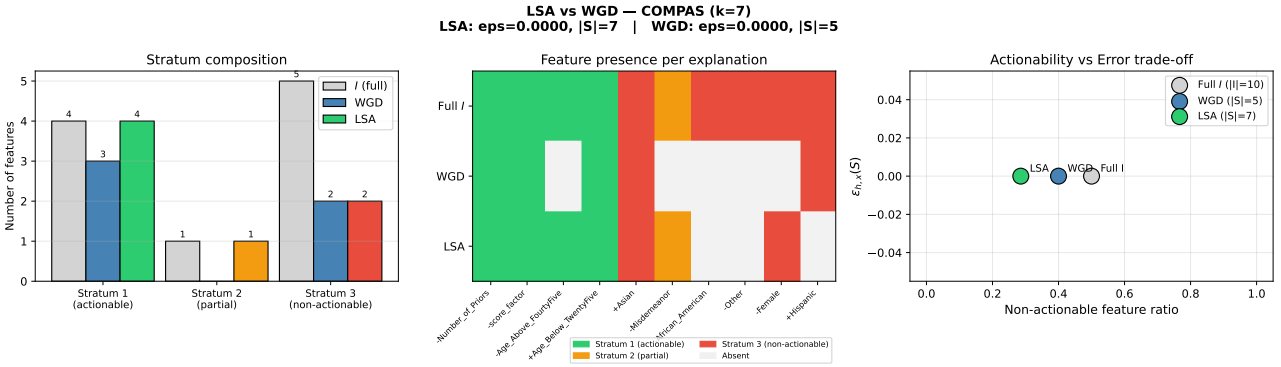


FIGURE 3 – LSA vs. WGD sur compas ($\epsilon_{max} = 0{,}1$, $k = 7$).

Plusieurs directions étendent naturellement ce travail. **(i)** L'extension des oracles efficaces pour $\mu_{h,x}$ aux forêts aléatoires, ensembles d'arbres et réseaux de neurones élargirait significativement l'applicabilité. **(ii)** Des bornes adaptatives à l'instance exploitant la structure du jeu de données ou de l'arbre pourraient donner des garanties plus serrées que le ratio de courbure dans le pire cas. **(iii)** Remplacer l'échantillonnage exhaustif de λ dans PFE par des stratégies bayésiennes permettrait d'identifier les points de coude plus efficacement. **(iv)** L'intégration de métriques d'équité de groupe comme contraintes dures dans WPPE renforcerait l'applicabilité dans les domaines réglementés tels que la justice pénale et le crédit. **(v)** Enfin, des études utilisateur réelles validant l'alignement des préférences au-delà des proxies synthétiques (SHAP, LIME, importance des caractéristiques) demeurent une étape essentielle vers le déploiement dans des systèmes de décision orientés vers l'humain.

Références

[1] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. In *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, Stockholm, Sweden, 2018. arXiv :1806.08049.

[2] Marcelo Arenas, Pablo Barceló, Miguel Romero Orth, and Bernardo Subercaseaux. On computing probabilistic explanations for decision trees. *Advances in Neural Information Processing Systems*, 35 :28695–28707, 2022.

[3] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J-M. Lagniez, and P. Marquis. Trading complexity for sparsity in random forest explanations. In *Proc. of AAAI'22*, 2022.

[4] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.M. Lagniez, and P. Marquis. On preferred abductive explanations for decision trees and random forests. In *Proc. of IJCAI'22*, 2022.

[5] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the computational intelligibility of boolean classifiers. In *Proc. of KR'21*, 2021.

[6] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the explanatory power of boolean decision trees. *Data Knowledge Engineering*, 142, 2022.

[7] Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. On the computation of example-based abductive explanations for random forests. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3679–3687, 2024.

[8] Steve Azzolin, SAGAR MALHOTRA, Andrea Passerini, and Stefano Teso. Beyond topological self-explainable GNNs : A formal explainability perspective. In *Forty-second International Conference on Machine Learning*, 2025.

[9] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. Model interpretability through the lens of computational complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 15487–15498, 2020.

[10] Shahaf Bassan, Xuanxiang Huang, and Guy Katz. Unifying formal explanations : A complexity-theoretic perspective. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. Poster.

[11] Shahaf Bassan and Guy Katz. Towards formal xai : Formally approximate minimal explanations of neural networks. In *Proceedings of the 29th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, 2022.

[12] Louenas Bounia and Frederic Koriche. Approximating probabilistic explanations via supermodular mi-

nimization. In *Uncertainty in Artificial Intelligence (UAI 2023).*, 2023.

[13] Louenas Bounia and Insaf Setitra. Enhancing the intelligibility of decision trees with concise and reliable probabilistic explanations. *Data & Knowledge Engineering*, 2025.

[14] Nelson Cowan. The magical number 4 in short-term memory : A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1) :87–114, 2001.

[15] Yves Crama and Peter L Hammer. *Boolean functions : Theory, algorithms, and applications*. Cambridge University Press, 2011.

[16] A. Darwiche and A. Hirth. On the reasons behind decisions. In *Proc. of ECAI'20*, 2020.

[17] Adnan Darwiche. Logic for explainable AI. In *Proceedings of the 38th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–11. IEEE, 2023.

[18] Indraneel Das and J. E. Dennis. A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems. *Structural Optimization*, 14(1) :63–69, 1997.

[19] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. In *Science Advances*, volume 4, page eaao5580, 2018.

[20] Matthias Ehrgott. *Multicriteria Optimization*. Springer-Verlag Berlin Heidelberg, 2nd edition, 2005.

[21] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639) :115–118, 2017.

[22] Xuanxiang Huang and João Marques-Silva. On the failings of shapley values for explainability. *Int. J. Approx. Reason.*, 2024.

[23] A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *Proc. of AAAI'19*, pages 1511–1519, 2019.

[24] Rishabh Iyer and Jeff Bilmes. Submodular optimization with submodular cover and submodular knapsack constraints. *NeurIPS*, 2013.

[25] Y. Izza and J. Marques-Silva. On explaining random forests with SAT. In *Proc. of IJCAI'21*, pages 2584–2591, 2021.

[26] Yacine Izza, Xuanxiang Huang, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and Joao Marques-Silva. On computing probabilistic abductive explanations. *Int. J. Approx. Reason.*, 2022.

[27] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On explaining decision trees. *ArXiv*, abs/2010.11034, 2020.

[28] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On tackling explanation redundancy in decision trees. *J. Artif. Intell. Res.*, 75 :261–321, 2022.

[29] Yacine Izza, Alexey Ignatiev, Sasha Rubin, Joao Marques-Silva, and Peter J. Stuckey. Most general explanations of tree ensembles (extended version). *IJCAI 2025*, 2025.

[30] Yacine Izza, Alexey Ignatiev, Peter J. Stuckey, and Joao Marques-Silva. Delivering inflated explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[31] Yacine Izza, Kuldeep S. Meel, and João Marques-Silva. Locally-minimal probabilistic explanations. *ECAI*, 2024.

[32] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse : Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5) :1–29, 2023.

[33] F. Koriche, J.-M. Lagniez, P. Marquis, and S. Thomas. Knowledge compilation for model counting : Affine decision trees. In *Proc. of IJCAI'13*, pages 947–953, 2013.

[34] Frédéric Koriche, Jean-Marie Lagniez, and Chi Tran. Probabilistic explanations for regression models. In *Proceedings on Uncertainty in Artificial Intelligence*, 2025.

[35] M. Suresh Kumar, V. Soundarya, S. Kavitha, E.S. Keerthika, and E. Aswini. Credit card fraud detection using random forest algorithm. In *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)*, 2019.

[36] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai) ? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296 :103473, 2021.

[37] S. Lundberg and S-I. Lee. A unified approach to interpreting model predictions. In *Proc. of NIPS'17*, 2017.

[38] Joao Marques-Silva. Logic-based explainability in machine learning. *ArXiv*, abs/2211.00541, 2023.

[39] Joao Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explaining naive bayes and other linear classifiers with polynomial time and delay. *Neural Information Processing Systems*, 2020.

[40] Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Boston, 1999.

[41] G. A. Miller. The magical number seven, plus or minus two : Some limits on our capacity for processing information. *The Psychological Review*, 63(2) :81–97, 1956.

[42] T. Miller. Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence*, 2019.

[43] C. Molnar. *Interpretable Machine Learning*. Leanpub, 2020.

[44] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617. ACM, 2020.

[45] George Nemhauser and Laurence Wolsey. *Integer and Combinatorial Optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., New York, 1988. Advisory Editors : Ronald L. Graham, Jan Karel Lenstra, Robert E. Tarjan.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

[47] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring : Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481, 2020.

[48] Yanou Ramon, Tom Vermeire, Olivier Toubia, David Martens, and Theodoros Evgeniou. Understanding consumer preferences for explanations generated by xai algorithms. *arXiv*, 2021.

[49] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you ?" : Explaining the predictions of any classifier. In *Proc. of SIGKDD'16*, pages 1135–1144, 2016.

[50] Cynthia Rudin. Please stop explaining black box models for high stakes decisions. *ArXiv*, abs/1811.10154, 2018.

[51] A. Shih, A. Choi, and A. Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proc. of IJCAI'18*, 2018.

[52] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap : Adversarial attacks on post hoc explanation methods. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 180–186. ACM, 2020.

[53] Kacper Sokol and Peter A. Flach. One explanation does not fit all : The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz*, 34(2) :235–250, 2020.

[54] Bernardo Subercaseaux, Marcelo Arenas, and Kuldeep S. Meel. Probabilistic explanations for linear models. In *Proceedings of the Thirty-Ninth AAAI Conference*, 2025.

[55] Maxim Sviridenko, Jan Vondrák, and Justin Ward. Optimal approximation for submodular and supermodular optimization with bounded curvature. *Mathematics of Operations Research*, 42(4) :1197–1218, 2017.

[56] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[57] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box : Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2) :841–887, 2018.

[58] Stephan Wäldchen. *Towards explainable artificial intelligence : Interpreting neural network classifiers with probabilistic prime implicants*. PhD thesis, Technische Universität Berlin, 2022. Ph.D. thesis.

[59] Eric Wang, Pasha Khosravi, and Guy Van den Broeck. Probabilistic sufficient explanations. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3082–3088, 2021.

[60] Min Wu, Haoze Wu, and Clark Barrett. Verix : Towards verified explainability of deep neural networks. In *NeurIPS*, 2023.

[61] Stephan Wäldchen, Jan MacDonald, Sascha Hauch, and Gitta Kutyniok. The computational complexity of understanding binary classifier decisions. *Journal of Artificial Intelligence Research*, 70 :351–387, 2021.

SESSION 5 : SAT, ASP présidée parFrédéric Maris

Sur le calcul des max-réfutations minimales

Meriam Ayari[1,2], Sami Cherif[1], Felip Manyà[2]

[1] Laboratoire MIS UR 4290, Université de Picardie Jules Verne, Amiens, France
[2] Artificial Intelligence Research Institute (IIIA), CSIC, Bellaterra, Spain

{meriam.ayari,sami.cherif}@u-picardie.fr felip@iiia.csic.es

Résumé

MaxSAT est une version d'optimisation de SAT qui consiste à déterminer le nombre maximum de clauses qu'il est possible de satisfaire par une affectation des variables. Les certificats d'optimalité pour les solutions MaxSAT sont essentiels pour la vérification et ont été principalement étudiés à travers le prisme des systèmes de preuve, la max-résolution étant la règle d'inférence la plus étudiée dans la littérature. Dans cet article, nous abordons le problème de la recherche de réfutations par max-résolution de taille minimale. Nous proposons une approche basée sur SAT, qui exploite les encodages propositionnels pour modéliser la recherche de réfutations par max-résolution de taille fixée, tout en respectant les contraintes de compensation et de consommation de clauses imposées par la règle.

Mots-clés

Satisfiabilité Maximum, Max-résolution, Réfutation

Abstract

MaxSAT is an optimization variant of SAT that aims to find an assignment maximizing the number of satisfied clauses. Certificates of optimality play a key role in the verification of MaxSAT solutions and are commonly studied through proof systems, with MaxSAT resolution being the most studied inference rule. In this paper, we study the problem of computing minimum-length refutations, measured by the number of inference steps, under the MaxSAT resolution rule. We propose a SAT-based approach that uses propositional encodings to model and search for MaxSAT resolution refutations of fixed size, while respecting structural, consumption, and compensation constraints imposed by the rule.

Keywords

Maximum Satisfiability, MaxSAT resolution, Refutation

1 Introduction

La satisfiabilité propositionnelle (SAT) et la satisfiabilité maximum (MaxSAT) sont deux formalismes puissants, avec une grande capacité d'expression en logique propositionnelle [4]. Étant donné une formule propositionnelle en forme normale conjonctive (FNC), le problème SAT consiste à déterminer s'il existe une affectation des variables qui satisfait toutes les clauses de la formule. MaxSAT est une extension naturelle de SAT sous forme de problème d'optimisation. Son objectif est de déterminer le nombre maximal de clauses pouvant être satisfaites par une affectation des variables.

L'importance pratique de MaxSAT a fortement augmenté ces dernières années. Ses applications couvrent plusieurs domaines, notamment la vérification matérielle et logicielle [13, 29], les problèmes de planification et d'ordonnancement [9, 30], ainsi que la bioinformatique [11, 12], parmi d'autres. Cependant, malgré ces avancées, MaxSAT reste fondamentalement plus difficile à résoudre que SAT, à la fois sur le plan théorique et pratique. Cette difficulté vient de la complexité plus élevée liée à la recherche de solutions et, surtout, à la preuve de leur optimalité.

L'étude des systèmes de preuve pour MaxSAT est devenue un sujet majeur pour mieux comprendre la complexité et les possibilités de résolution efficace des problèmes d'optimisation en logique propositionnelle. Un système de preuve clausal pour MaxSAT est généralement défini par un ensemble de règles d'inférence, dont les prémisses et les conclusions sont exprimées sous forme d'ensembles de clauses. Dans ce cadre, une preuve d'optimalité, souvent appelée certificat d'optimalité, fournit une garantie formelle que le coût retourné par un solveur MaxSAT correspond effectivement à l'optimum réel de la formule donnée en entrée.

Bien que les réfutations fondées sur la résolution sont bien établies pour SAT, l'étude systématique et la génération pratique de tels certificats dans le cadre de MaxSAT ont reçu relativement peu d'attention jusqu'à récemment [2, 26, 27, 31, 32]. Dans ce contexte, la max-résolution [6], qui généralise la règle de résolution classique pour SAT [28], est l'un des systèmes de preuve les plus étudiés pour MaxSAT. En particulier, cette règle a été largement utilisée dans la résolution effective de MaxSAT, notamment dans les algorithmes de type séparation et évaluation (*Branch and Bound*) pour MaxSAT [7, 20, 21], ainsi que dans les approches MaxSAT basées sur des appels itératifs à des oracles SAT [14, 24]. Des avancées récentes ont également montré que la max-résolution, lorsqu'elle est complétée par des règles d'inférence simples, peut produire des preuves de taille polynomiale pour le principe des tiroirs [18, 5].

Cependant, il existe des différences fondamentales entre la max-résolution et la résolution classique utilisée dans SAT. En effet, dans SAT, l'insatisfiabilité peut être démontrée

par une réfutation composée d'une suite d'étapes de résolution qui permettent de dériver la clause vide. La règle de résolution classique permet à plusieurs étapes d'inférence de partager les mêmes clauses prémisses, car ces clauses restent dans la formule après chaque inférence. En revanche, les suites d'étapes de max-résolution sont beaucoup plus contraintes. En effet, la max-résolution remplace les clauses prémisses par les conclusions dérivées. Cette différence de fonctionnement est souvent décrite comme un « mouvement de connaissances ».[17], modifie profondément la structure et les propriétés des dérivations dans les preuves pour MaxSAT. De plus, afin de garantir la correction du raisonnement, la construction de preuves par max-résolution nécessite des mécanismes de compensation supplémentaires, mis en place par l'introduction de clauses de compensation. Ces exigences mettent en évidence un écart fondamental dans notre compréhension de la relation entre les systèmes d'inférence SAT et MaxSAT [5, 8, 25].

Des travaux récents dans [23] ont proposé une approche basée sur SAT pour calculer les plus courtes preuves de résolution dans le cadre de SAT. Dans cet article, nous lançons une étude préliminaire visant à étendre cette approche au cadre de la max-résolution. En nous appuyant sur le cadre proposé dans [23], nous adaptons l'encodage basé sur SAT afin de prendre en compte les contraintes supplémentaires et les propriétés structurelles propres à la max-résolution. En particulier, nous proposons un cadre pour calculer les plus courtes réfutations sous la règle de max-résolution, en encodant la recherche de réfutations MaxSAT de taille fixée sous forme de problèmes SAT. Notre objectif est d'établir des bornes précises sur la longueur des preuves d'optimalité pour les instances MaxSAT. Cela permet de combler une lacune identifiée dans la littérature et de construire des certificats plus courts et plus faciles à interpréter. Plus largement, ce travail contribue à une meilleure compréhension de la complexité des preuves dans les variantes d'optimisation de la satisfiabilité. Il soutient également le développement de certificats d'optimalité plus courts et plus efficaces pour les solveurs MaxSAT.

Le reste de cet article est organisé comme suit. La section 2 introduit le problème (Max-)SAT et présente la règle de max-résolution. La section 3 présente notre méthodologie pour la recherche de réfutations minimales. La section 4 détaille l'encodage SAT utilisé pour calculer des réfutations par max-résolution de longueur fixée. La section 5 présente nos expérimentations préliminaires ; enfin, nous concluons et discutons des perspectives de recherche dans la section 6.

2 Preliminaries

2.1 Définitions et notations

Soit $X = \{x_1, x_2, \ldots, x_n\}$ l'ensemble des variables propositionnelles. Un littéral l est soit une variable $x \in X$, soit sa négation $\neg x$ également notée $\overline{x}$. Une *clause* $c = (l_1 \vee l_2 \vee \cdots \vee l_k)$ est une disjonction de k littéraux, et peut aussi être représentée comme un ensemble de littéraux. Une *formule en forme normale conjonctive (FNC)* $\phi = c_1 \wedge c_2 \wedge \cdots \wedge c_m$ est une conjonction de m clauses, et peut être représentée comme un ensemble de clauses. Une *affectation* $\alpha : X \to \{0, 1\}$ associe chaque variable à une valeur booléenne et peut être représentée comme un ensemble de littéraux satisfaits. Un littéral l est *satisfait* par une affectation α si $l \in \alpha$, et il est *falsifié* si $\neg l \in \alpha$. Une clause c est *satisfaite* par une affectation α si au moins l'un de ses littéraux est satisfait par α ; sinon, elle est *falsifiée* par α. La *clause vide*, notée $\square$, ne contient aucun littéral et ne peut être satisfaite par aucune affectation. Deux clauses c et c' sont dites *opposées* s'il existe un littéral $l \in c$ tel que $\neg l \in c'$. Nous notons par $var(c)$ l'ensemble des variables qui apparaissent dans la clause c. Une formule FNC ϕ est *satisfaite* par une affectation α appelée *modèle* de ϕ, si toutes les clauses de ϕ sont satisfaites par α ; sinon, ϕ est *falsifiée*. Une affectation α est dite *complète* si chaque variable de ϕ reçoit une valeur ; sinon, elle est dite *partielle*. Le problème de satisfiabilité (SAT) consiste à déterminer s'il existe une affectation des variables qui satisfait une formule FNC donnée ϕ. Si une telle affectation existe, ϕ est dite *satisfiable* ; sinon, elle est dite *insatisfiable*.

Le problème de satisfiabilité maximum (MaxSAT) est une extension naturelle de SAT sous forme de problème d'optimisation. Il consiste simplement, étant donnée une formule FNC ϕ, à déterminer le nombre maximal de clauses qui peuvent être satisfaites par une affectation des variables. Une définition plus courante dans la littérature consiste à déterminer le nombre minimal de clauses que toute affectation de ϕ doit falsifier ; ce nombre est appelé l'optimum de ϕ. Formellement, soit $cost_\alpha(\phi)$ le nombre de clauses falsifiées par l'affectation α dans ϕ. L'*optimum* de ϕ est défini par $opt(\phi) = \min_\alpha cost_\alpha(\phi)$. Notons que si une formule FNC ϕ est satisfiable, alors $opt(\phi) = 0$; sinon, $opt(\phi) \geq 1$, car au moins une clause est falsifiée par chaque affectation de ϕ.

2.2 Réfutations par max-résolution

Pour certifier qu'une formule FNC est satisfiable, il suffit de donner un modèle de cette formule. Cependant, pour prouver qu'une formule est insatisfiable, il faut réfuter l'existence de tout modèle. Le système de réfutation le plus connu pour SAT est basé sur la règle de *résolution* [10]. Cette règle permet de dériver une clause appelée *résolvante* à partir de deux clauses opposées sur un littéral, comme défini ci-dessous. Cette résolvante peut être ajoutée à la formule tout en préservant sa satisfaisabilité. Ainsi, une *réfutation par résolution* d'une formule insatisfiable donnée ϕ est une suite finie d'étapes de résolution qui commence avec des clauses de ϕ et qui finit par dériver la clause vide $\square$. Les réfutations par résolution peuvent être représentées sous forme de *graphe orienté acyclique (DAG)*, où les nœuds représentent les clauses et les arcs représentent les dépendances entre les résolvantes et leurs prémisses.

Definition 1 (Résolution [10])**.** *Étant donné deux clauses $c_1 = (x \vee A)$ et $c_2 = (\neg x \vee B)$ opposées sur la variable x, où A et B sont des disjonctions de littéraux, la règle de résolution permet de dériver une clause c_r appelée la*

résolvante, comme suit :

$$\frac{c_1 = (x \vee A) \quad c_2 = (\neg x \vee B)}{c_r = (A \vee B)}$$

Contrairement à SAT, où les systèmes de preuve cherchent à démontrer l'insatisfiabilité, les systèmes de preuve pour MaxSAT ont un objectif différent. Ils permettent de calculer l'*optimum* d'une formule, c'est-à-dire le nombre minimal de clauses falsifiées parmi toutes les affectations possibles. Plusieurs systèmes de preuve complets pour MaxSAT ont été proposés dans la littérature, notamment la max-résolution [28] et les tableaux sémantiques [22]. Dans ces systèmes, la formule est réfutée un nombre de fois égal à son optimum, au moyen de transformations qui préservent l'équivalence, comme défini ci-dessous.

Definition 2 (Équivalence MaxSAT)**.** *Deux formules FNC ϕ et ϕ' sont équivalentes au sens de MaxSAT si, pour toute affectation $\alpha : var(\phi) \cup var(\phi') \rightarrow \{0,1\}$, nous avons $cost_\alpha(\phi) = cost_\alpha(\phi')$.*

L'un des premiers systèmes complets, et aussi l'un des plus étudiés pour MaxSAT, est basé sur la règle de max-résolution. Cette règle étend la règle de résolution classique de SAT au cadre de MaxSAT. La max-résolution, définie formellement ci-dessous, prend comme prémisses un ensemble de clauses opposées et déduit une résolvante, comme dans la résolution classique. Cependant, deux différences fondamentales entre la max-résolution et la résolution classique apparaissent afin de garantir la correction de la règle, au sens de l'équivalence MaxSAT. Premièrement, d'autres clauses sont dérivées en même temps que la résolvante. Ces clauses sont appelées clauses de compensation et doivent être ajoutées à la formule après l'application de la règle afin de préserver l'équivalence. Deuxièmement, contrairement à la résolution classique, où les prémisses sont conservées après chaque étape d'inférence, la max-résolution consomme ses prémisses. Autrement dit, une fois les clauses prémisses résolues, elles sont remplacées par les conclusions dérivées, formées par la résolvante et les clauses de compensation. Elles ne peuvent donc plus être utilisées comme prémisses dans des dérivations ultérieures sous leur forme initiale.

Definition 3 (Max-résolution [16])**.** *Étant donné deux clauses $C_1 = x \vee A$ et $C_2 = \neg x \vee B$ opposées sur la variable x, où $A = a_1 \vee \cdots \vee a_s$ et $B = b_1 \vee \cdots \vee b_t$, la règle de max-résolution est définie comme suit :*

$$\frac{C_1 = x \vee A \qquad C_2 = \neg x \vee B}{C_r = A \vee B}$$

$$\begin{aligned} CC_1 &= x \vee A \vee \neg b_1 \\ CC_2 &= x \vee A \vee b_1 \vee \neg b_2 \\ &\vdots \\ CC_t &= x \vee A \vee b_1 \vee \cdots \vee b_{t-1} \vee \neg b_t \end{aligned}$$

$$\begin{aligned} CC_{t+1} &= \neg x \vee B \vee \neg a_1 \\ CC_{t+2} &= \neg x \vee B \vee a_1 \vee \neg a_2 \\ &\vdots \\ CC_{t+s} &= \neg x \vee B \vee a_1 \vee \cdots \vee a_{s-1} \vee \neg a_s \end{aligned}$$

Notons que la max-résolution, telle qu'elle est présentée dans la définition précédente, dépend de l'ordre des littéraux. Autrement dit, selon l'ordre des littéraux dans les clauses prémisses, les clauses de compensation dérivées peuvent avoir des formes différentes. Cependant, toutes ces formes sont équivalentes. Nous pouvons donc les représenter de manière abstraite à l'aide d'une écriture plus compacte de la règle, comme suit :

$$\frac{c_1 = (x \vee A) \qquad c_2 = (\neg x \vee B)}{\begin{array}{ll} c_r = A \vee B & \text{(Résolvante)} \\ CC_1 \equiv x \vee A \vee \overline{B} & \text{(Ensemble de compensation 1)} \\ CC_2 \equiv \neg x \vee B \vee \overline{A} & \text{(Ensemble de compensation 2)} \end{array}}$$

où $\overline{A} = (\overline{a_1}) \wedge (a_1 \vee \overline{a_2}) \wedge \cdots \wedge (a_1 \vee \cdots \vee a_{s-1} \vee \overline{a_s})$, et de manière similaire pour B.

Une *réfutation par max-résolution*, ou simplement *max-réfutation*, est une suite finie d'étapes de max-résolution qui commence avec les clauses de ϕ et dérive une formule ϕ' telle que $\square \in \phi'$. De telles preuves peuvent aussi être représentées sous forme de DAG, où les nœuds représentent les clauses et les arcs représentent les dépendances d'inférence. La *taille* d'une réfutation est le nombre de ses étapes d'inférence, c'est-à-dire le nombre total de clauses résolvantes générées par la max-résolution.

Example 1. *Considérons la formule FNC $\phi = \{(\neg x_1 \vee x_3), (x_1), (\neg x_1 \vee x_2), (\neg x_2 \vee \neg x_3)\}$. La figure 1 illustre deux max-réfutations valides issues de [8]. Elles se distinguent par l'ordre dans lequel sont effectuées les étapes de résolution impliquant la clause non-read-once x_1, avec des tailles égales à 4 pour la dérivation de gauche et à 5 pour celle de droite.*

3 Méthodologie

Dans cette section, nous nous intéressons au problème de la recherche des plus courtes réfutations par max-résolution. Cette tâche est fondamentalement plus difficile que dans le cadre de la résolution classique étudié dans [23], en raison du caractère consommant de la max-résolution, de la présence de contraintes de compensation et de sa sensibilité à l'ordre des étapes d'inférence. À ce stade, nous nous limitons aux plus courtes réfutations, plutôt qu'aux plus courtes preuves d'optimalité de MaxSAT. Néanmoins, notre approche ouvre des pistes prometteuses pour traiter ce dernier problème. Contrairement à la résolution classique, où les prémisses peuvent être réutilisées et où l'ordre des littéraux n'a pas d'importance, la max-résolution impose des contraintes strictes sur ces deux aspects. Cela rend la recherche systématique de preuves de longueur minimale non triviale. De manière similaire à [23], nous abordons le problème du calcul des plus courtes réfutations par max-résolution en développant une méthodologie d'encodage basée sur SAT. Nous expliquons maintenant l'idée et les hypothèses centrales avant de présenter l'encodage formel. Tout d'abord, on veut décider si une formule FNC insatisfiable donnée F admet une réfutation par max-résolution

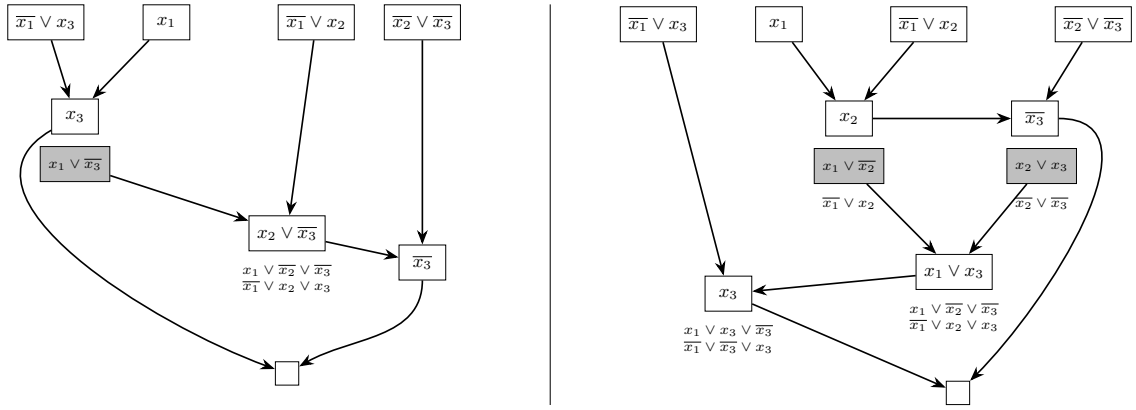


FIGURE 1 – Deux réfutations possibles par max-résolution pour une même formule. Les clauses utilisées sont indiquées dans des boîtes, et les clauses de compensation utilisées sont surlignées en gris.

Algorithm 1 Max-réfutation de longueur minimale

Require: Formule insatisfiable F
1: $K \leftarrow 1$
2: **loop**
3: $\Phi_K \equiv \exists$ une max-réfutation de F de taille K ?
4: (résultat, modèle) $\leftarrow$ SOLVE(Φ_K)
5: **if** résultat = SAT **then**
6: **return** (K, modèle)
7: **else**
8: $K \leftarrow K + 1$
9: **end if**
10: **end loop**

avec exactement K étapes d'inférence. Pour cela, nous construisons une formule propositionnelle Φ_K telle que :

— si Φ_K est satisfiable, alors on peut extraire, à partir d'une affectation satisfaisante, une réfutation par max-résolution de F de taille K ;
— si Φ_K est insatisfiable, alors aucune telle réfutation n'existe dans l'abstraction capturée par l'encodage.

Comme la profondeur d'une plus courte réfutation n'est pas connue à l'avance, nous adoptons une approche itérative sur la longueur de preuve K, décrite dans l'algorithme 1. À chaque itération, nous encodons la formule Φ_K qui représente l'existence d'une max-réfutation de F de taille fixée K. Si le solveur retourne $UNSAT$, alors K constitue naturellement une borne inférieure stricte sur la taille minimale de la réfutation. Nous augmentons donc K de manière itérative, jusqu'à ce que le solveur retourne SAT. Dans ce cas, nous obtenons la taille optimale de la preuve, et nous pouvons également extraire la preuve calculée à partir de l'affectation retournée par le solveur.

Comme la max-résolution est plus complexe que la résolution classique, nous adoptons des des choix pour des fins de simplification, qui sont décrits dans la suite. Rappelons que la max-résolution est un système de preuve destructif : une fois qu'une clause est utilisée comme prémisse, elle est consommée et ne peut plus être réutilisée, sous sa forme inchangée, plus tard dans la dérivation. Pour représenter fidèlement ce comportement, notre encodage doit modéliser explicitement la disponibilité temporelle des clauses ainsi que leur consommation. De plus, à chaque étape d'inférence, une application de la max-résolution produit une résolvante et plusieurs clauses de compensation dérivées de chaque parent. La principale difficulté consiste donc à représenter ce qui reste disponible après chaque étape d'inférence. Les clauses standards sont simples à représenter, car elles sont soit encore disponibles, soit déjà consommées. Les clauses de compensation sont plus difficiles à gérer, car elles forment des ensembles structurés dont l'utilisation future dépend des littéraux déjà fixés par les étapes précédentes et des littéraux qui peuvent encore participer à des résolutions ultérieures. Pour capturer cela sans énumérer explicitement tous les descendants possibles des clauses de compensation, nous représentons chaque ensemble de compensation, comme dans la forme condensée de la max-résolution, par une paire (F, R), où :

— F est la *partie fixe*, contenant les littéraux déjà engagés et qui persistent dans toute clause future extraite de cet ensemble de compensation ;
— R est la *partie restante*, contenant les littéraux encore actifs et pouvant être consommés plus tard.

Intuitivement, F enregistre l'historique irréversible de l'ensemble de compensation, tandis que R représente son potentiel restant pour la recherche de preuve. Cette abstraction traite donc les clauses de compensation sous la forme $F \vee \overline{R}$, constituant ainsi le choix de modélisation central de notre cadre. En toute généralité, utiliser un ensemble de compensation comme parent peut le diviser en plusieurs descendants, selon la position du pivot et la clause extraite de l'ensemble. Encoder directement toutes ces possibilités conduirait à une forte explosion combinatoire. Pour cette raison, nous limitons notre attention à trois schémas de transition canoniques, suffisants pour notre procédure de recherche de preuve. L'un de ces cas repose en plus sur une *restriction de plus grande clause*, afin d'éviter qu'un parent de compensation ne génère plusieurs descendants simultanément. Cette restriction simplifie l'encodage, mais elle implique aussi que notre cadre calcule les plus courtes réfutations

dans le système de transition restreint que nous encodons. Plus précisément, lorsqu'un ensemble de compensation (F, R) est sélectionné comme parent à une étape d'inférence, la clause extraite doit contenir le pivot avec la bonne polarité pour le côté choisi de l'inférence : à gauche, la clause extraite doit contenir x, tandis qu'à droite elle doit contenir $\neg x$. La transition dépend alors de la manière dont ce pivot est obtenu à partir de l'ensemble de compensation. Pour éviter toute ambiguïté, nous adoptons la convention suivante dans cette section : x désigne toujours la variable pivot sélectionnée à l'étape de résolution courante, tandis que y désigne un littéral non pivot choisi dans la partie restante R lorsqu'un tel choix est nécessaire. De plus, sauf mention contraire, nous supposons que le parent gauche fournit le pivot positivement et que le parent droit le fournit négativement. Sous cette convention, nous distinguons trois cas canoniques, selon la position du pivot dans la structure de compensation.

Cas 1 : le pivot est déjà dans la partie fixe. Supposons d'abord que l'ensemble de compensation soit utilisé comme parent gauche. La clause extraite doit donc contenir le pivot positif x. Si $x \in F$, alors le pivot apparaît déjà dans toutes les clauses représentées par l'objet de compensation. Dans ce cas, il n'est pas nécessaire de consommer le pivot depuis la partie restante. À la place, nous consommons un autre littéral $y \in R$, avec $y \neq x$, et nous extrayons la clause

$$F \vee x \vee \neg y.$$

L'ensemble de compensation mis à jour (F', R') est donné par

$$F' = F, \qquad R' = R \setminus \{y\}.$$

Intuitivement, le pivot est hérité de la partie fixe. La transition réduit donc seulement la partie restante. C'est le cas le plus simple, car le pivot ne modifie pas la structure interne de l'ensemble de compensation.

Example 2. *Soient $F = \{x, a\}$ et $R = \{b, c\}$, et supposons que le parent gauche doive fournir le pivot x. Comme $x \in F$, nous pouvons choisir $y = b$ et extraire la clause $x \vee a \vee \neg b$. L'ensemble de compensation obtenu est $(\{x, a\}, \{c\})$.*

Cas 2 : le pivot est dans la partie restante avec la polarité opposée. Supposons encore que l'ensemble de compensation soit utilisé comme parent gauche et qu'il doit fournir le pivot positif x. Si $x \notin F$ mais $\neg x \in R$ sous une forme qui permet une extraction directe, alors le pivot est obtenu en consommant ce littéral depuis la partie restante. La clause extraite est

$$F \vee x.$$

L'ensemble de compensation mis à jour (F', R') est alors donné par

$$F' = F \cup \{\neg x\}, \qquad R' = R \setminus \{\neg x\}.$$

Intuitivement, le littéral pris dans la partie restante est consommé et devient une partie de l'historique fixe avec la polarité opposée. Cela reflète le fait qu'une fois que l'ensemble de compensation a été utilisé pour fournir le pivot à partir de la partie restante, ce choix devient irréversible.

Example 3. *Soient $F = \{a\}$ et $R = \{\neg x, b\}$, et supposons que le parent gauche doit fournir le pivot x. Nous extrayons la clause $a \vee x$. L'ensemble de compensation mis à jour devient $(\{a, \neg x\}, \{b\})$.*

Cas 3 : le pivot est dans la partie restante avec la même polarité. La situation la plus délicate apparaît lorsque l'ensemble de compensation est utilisé comme parent gauche, que la clause extraite doit contenir x, et que x n'est pas déjà disponible dans la partie fixe, mais apparaît avec la même polarité dans la partie restante, c'est-à-dire $x \in R$. Extraire x de la partie restante peut alors nécessiter de diviser l'ensemble de compensation en plusieurs descendants. Pour garder l'encodage compact, nous ne représentons pas ce cas dans toute sa généralité. À la place, nous limitons la transition à la forme de *plus grande clause*

$$F \vee x \vee (R \setminus \{y\}) \vee \neg y,$$

pour un certain littéral choisi $y \in R$, puis nous mettons à jour l'ensemble de compensation comme suit :

$$F' = F, \qquad R' = R \setminus \{y\}.$$

Intuitivement, ce cas simule l'effet d'une consommation dans la partie restante sans générer explicitement plusieurs objets de compensation successifs. La restriction de la plus grande clause n'est donc pas une propriété intrinsèque de la max-résolution, mais plutôt une restriction de notre cadre d'encodage introduite pour contrôler l'espace de recherche.

Example 4. *Soient $F = \{a\}$ et $R = \{x, b, c\}$, et supposons que le parent gauche doit fournir le pivot x. Au lieu de diviser la structure de compensation, nous choisissons $y = b$ et nous extrayons la clause*

$$a \vee x \vee c \vee \neg b.$$

L'ensemble de compensation mis à jour est donc $(\{a\}, \{x, c\})$.

Supposons maintenant que nous extrayions plutôt la clause $a \vee x \vee \neg b$, c'est-à-dire sans inclure les autres littéraux restants de R. Cette clause correspond à l'utilisation d'un préfixe partiel de la partie restante. Dans ce cas, les clauses restantes, c'est-à-dire $a \vee \neg x$ et $a \vee x \vee c \vee \neg b$, ne peuvent pas être représentées par un seul ensemble de compensation successeur : il faut distinguer les clauses qui incluent c de celles qui ne l'incluent pas. Cela conduit à une division *de l'ensemble de compensation en plusieurs descendants, chacun représentant un sous-ensemble différent des littéraux restants. En revanche, l'extraction de la plus grande clause détermine de manière unique la structure résiduelle, qui peut être représentée par un seul ensemble de compensation.*

Nous remarquons que les cas 1 et 2 ne nécessitent jamais de division : dans le cas 1, le pivot est déjà présent dans la partie fixe et donc disponible globalement, tandis que dans le cas 2, le pivot complémentaire est directement retiré de la partie restante et transféré vers la partie fixe. Seul le cas 3, où x apparaît positivement dans la partie restante R sans être déjà fixé, introduit un choix de branchement sur la quantité de R à déplier avant d'utiliser x. La restriction de la plus grande clause résout cette ambiguïté en sélectionnant une seule option canonique, ce qui garantit que chaque parent de compensation produit exactement un seul descendant de son côté. Les deux premiers cas correspondent directement à des manières naturelles d'extraire le pivot de la partie fixe et la partie restante. Le troisième cas est une restriction volontaire utilisée pour éviter un branchement sur plusieurs descendants de compensation. Par conséquent, la minimalité dans notre cadre est régie par la sémantique de transition de la forme de plus grande clause, ce qui peut entraîner un surcoût dans la taille des preuves générées.

4 Modélisation formelle des max-réfutations de longueur fixée

Avant de présenter les contraintes une par une, nous donnons d'abord une vue d'ensemble de l'encodage. Un modèle de Φ_K représente une réfutation par max-résolution de taille exactement K, dans le système de transition restreint de la section 3. Plus précisément, pour chaque étape $t \in \{1, \ldots, K\}$, le modèle spécifie : (i) une variable pivot, (ii) un parent gauche et un parent droit, chacun pouvant être soit une clause standard, soit un ensemble de compensation, (iii) le contenu de la résolvante produite à l'étape t, (iv) les objets de compensation générés à cette étape, et (v) l'état de disponibilité de tous les objets pour les étapes futures. Pour capturer ces informations, nous introduisons d'abord plusieurs familles de variables, puis nous organisons les contraintes en couches sémantiques, afin de garantir que chaque étape correspond à une inférence valide.

4.1 Univers et variables

Afin de garantir un indexage correct, nous utilisons les univers suivants :

- $\mathcal{X}$: ensemble des variables de la formule.
- $\mathcal{S} = \{1, \ldots, K\}$: indices des étapes d'inférence.
- $\mathcal{D} = \{1, \ldots, M + K\}$: indices des clauses *standards*, où $\mathcal{I} = \{1, \ldots, M\}$ sont les indices des clauses initiales de la formule d'entrée, et $\mathcal{J} = \{M + 1, \ldots, M + K\}$ sont les indices des résolvantes générées pendant la dérivation.
- $\mathcal{C} = \{1, \ldots, 2K\}$: indices des ensembles de compensation, où les ensembles générés à l'étape t sont indexés par $2t-1$ pour l'ensemble de compensation gauche, et par $2t$ pour l'ensemble de compensation droit.

À l'étape t, seuls les objets créés strictement avant t peuvent être sélectionnés comme parents. Nous définissons donc les ensembles disponibles suivants afin de garantir la cohérence temporelle :

$$\mathcal{D}^t = \{i \in \mathcal{D} \mid i < M + t\}$$

$$\mathcal{C}^t = \{k \in \mathcal{C} \mid k < 2t - 1\}.$$

Nous pouvons maintenant définir les variables propositionnelles utilisées dans l'encodage :

Variables de sélection.

- $v_{x,t}$: la variable $x \in \mathcal{X}$ est sélectionnée comme pivot à l'étape $t \in \mathcal{S}$.
- $ls_{t,i}, rs_{t,i}$: à l'étape t, la clause standard $i \in \mathcal{D}^t$ est sélectionnée comme parent gauche/droit.
- $lcc_{t,k}, rcc_{t,k}$: à l'étape t, l'ensemble de compensation $k \in \mathcal{C}^t$ est sélectionné comme parent gauche/droit.

Variables de contenu.

- $ps_{x,i}, ns_{x,i}$: le littéral x, respectivement $\neg x$, apparaît dans la clause standard $i \in \mathcal{D}$.
- $pccF_{x,k,t}, nccF_{x,k,t}$: le littéral x, respectivement $\neg x$, appartient à la partie fixe de l'ensemble de compensation k à l'étape t.
- $pccR_{x,k,t}, nccR_{x,k,t}$: le littéral x, respectivement $\neg x$, appartient à la partie restante de l'ensemble de compensation k à l'étape t.

Variables de disponibilité.

- $as_{t,i}$: la clause standard i est disponible à l'étape t.
- $acc_{t,k}$: l'ensemble de compensation k est disponible à l'étape t.
- $accx_{x,t,k}$: à l'étape t, un littéral sur la variable x est consommé depuis la partie restante de l'ensemble de compensation k.

4.2 Sémantique de sélection

Nous imposons d'abord la validité structurelle de la dérivation. À chaque étape, exactement une variable pivot est sélectionnée, avec exactement un parent gauche et exactement un parent droit. Les parents peuvent être soit des clauses standards, soit des ensembles de compensation. De plus, le même objet ne peut pas être sélectionné des deux côtés de l'inférence.

$$\forall t \in \mathcal{S} :$$

$$\sum_{x \in \mathcal{X}} v_{x,t} = 1 \tag{1}$$

$$\sum_{i \in \mathcal{D}^t} ls_{t,i} + \sum_{k \in \mathcal{C}^t} lcc_{t,k} = 1 \tag{2}$$

$$\sum_{i \in \mathcal{D}^t} rs_{t,i} + \sum_{k \in \mathcal{C}^t} rcc_{t,k} = 1 \tag{3}$$

$$\overline{ls_{t,i}} \vee \overline{rs_{t,i}} \qquad \forall i \in \mathcal{D}^t \tag{4}$$

$$\overline{lcc_{t,k}} \vee \overline{rcc_{t,k}} \qquad \forall k \in \mathcal{C}^t \tag{5}$$

Pour chaque parent de compensation sélectionné, au plus une variable peut être consommée depuis sa partie restante. Cela impose qu'un objet de compensation évolue selon une seule transition à chaque étape d'inférence.

$$\forall t \in \mathcal{S}, \forall k \in \mathcal{C}^t : \sum_{x \in \mathcal{X}} accx_{x,t,k} \leq 1. \quad (6)$$

Enfin, un objet ne peut être sélectionné comme parent que s'il est disponible à l'étape courante. Cette condition s'applique à la fois aux clauses standards et aux ensembles de compensation.

$$(ls_{t,i} \vee rs_{t,i}) \implies as_{t,i} \quad \forall t \in \mathcal{S}, \forall i \in \mathcal{D}^t \quad (7)$$
$$(lcc_{t,k} \vee rcc_{t,k}) \implies acc_{t,k} \quad \forall t \in \mathcal{S}, \forall k \in \mathcal{C}^t \quad (8)$$

Les contraintes (1)–(8) garantissent la cohérence structurelle de chaque étape d'inférence. La contrainte (1) sélectionne exactement une variable pivot. Les contraintes (2) et (3) sélectionnent respectivement un parent gauche et un parent droit. Les contraintes (4) et (5) empêchent la même clause standard ou le même ensemble de compensation d'être sélectionné des deux côtés de la même inférence. La contrainte (6) garantit qu'un parent de compensation sélectionné consomme au plus une variable depuis sa partie restante. Enfin, les contraintes (7) et (8) garantissent que seuls les objets disponibles peuvent être sélectionnés comme parents.

4.3 Sémantique FNC

Nous encodons maintenant le contenu des clauses standards et assurons la cohérences des objets traités. D'abord, pour chaque clause initiale $i \in \mathcal{I}$ et chaque variable $x \in \mathcal{X}$, la valeur de vérité des variables de contenu correspondantes est fixée par la formule d'entrée.

$$ps_{x,i} \quad \text{si } x \in C_i, \quad \overline{ps_{x,i}} \quad \text{sinon,} \quad (9)$$
$$ns_{x,i} \quad \text{si } \neg x \in C_i, \quad \overline{ns_{x,i}} \quad \text{sinon.} \quad (10)$$

Ensuite, nous imposons la cohérence des polarités : aucune variable ne peut apparaître avec les deux polarités dans la même clause standard, dans la même partie fixe ou dans la même partie restante.

$$\overline{ps_{x,i}} \vee \overline{ns_{x,i}} \quad \forall i \in \mathcal{J}, \forall x \in \mathcal{X} \quad (11)$$
$$\overline{pccF_{x,k,t}} \vee \overline{nccF_{x,k,t}} \quad \forall t \in \mathcal{S}, \forall k \in \mathcal{C}, \forall x \in \mathcal{X} \quad (12)$$
$$\overline{pccR_{x,k,t}} \vee \overline{nccR_{x,k,t}} \quad \forall t \in \mathcal{S}, \forall k \in \mathcal{C}, \forall x \in \mathcal{X} \quad (13)$$

Enfin, nous imposons la séparation entre la partie fixe et la partie restante. Autrement dit, un littéral peut appartenir soit à la partie fixe, soit à la partie restante d'un ensemble de compensation, mais jamais aux deux.

$$\forall x \in \mathcal{X}, \forall t \in \mathcal{S}, \forall k \in \mathcal{C} :$$
$$\overline{pccF_{x,k,t}} \vee \overline{pccR_{x,k,t}} \quad (14)$$
$$\overline{pccF_{x,k,t}} \vee \overline{nccR_{x,k,t}} \quad (15)$$
$$\overline{nccF_{x,k,t}} \vee \overline{pccR_{x,k,t}} \quad (16)$$
$$\overline{nccF_{x,k,t}} \vee \overline{nccR_{x,k,t}} \quad (17)$$

Les contraintes (9) et (10) initialisent le contenu des clauses d'entrée selon la formule originale. Les contraintes (11)–(13) imposent une cohérence locale des polarités, en garantissant qu'aucune variable n'apparaît avec les deux polarités dans une même clause standard, une même partie fixe ou une même partie restante. Enfin, les contraintes (14)–(17) garantissent que les parties fixe et restante de chaque ensemble de compensation sont disjointes.

4.4 Sémantique d'inférence

Dans cette section, nous imposons la sémantique de la max-résolution. Par convention, dans la suite de l'encodage, la polarité du pivot est fixée comme suit : le parent gauche doit contenir la variable pivot x avec une polarité positive, tandis que le parent droit doit la contenir avec une polarité négative. Pour les clauses standards, cette exigence est directe : si une clause standard est sélectionnée comme parent gauche, elle doit contenir x ; si elle est sélectionnée comme parent droit, elle doit contenir $\neg x$.

$$\forall t \in \mathcal{S}, \forall x \in \mathcal{X}, \forall i \in \mathcal{D}^t :$$
$$(ls_{t,i} \wedge v_{x,t}) \implies ps_{x,i} \quad (18)$$
$$(rs_{t,i} \wedge v_{x,t}) \implies ns_{x,i} \quad (19)$$

Pour les ensembles de compensation, la situation est plus subtile, car un objet de compensation (F, R) représente une famille de clauses plutôt qu'une seule clause. Ainsi, les contraintes suivantes ne doivent pas être interprétées comme imposant que plusieurs littéraux apparaissent simultanément dans une même clause concrète. Elles expriment plutôt une *condition d'admissibilité* : un ensemble de compensation peut servir de parent pour le pivot x lorsqu'il contient assez d'information pour extraire une clause concrète où le pivot apparaît avec la polarité requise par le côté sélectionné.

Plus précisément, le pivot peut être justifié par l'un des trois cas de transition introduits dans la section 3 :

- le pivot est déjà présent dans la partie fixe ($pccF_{x,k,t-1}$ pour le parent gauche, $nccF_{x,k,t-1}$ pour le parent droit), ce qui correspond au cas 1 ;
- le pivot apparaît dans la partie restante avec la polarité requise par le côté sélectionné ($pccR_{x,k,t-1}$ pour le parent gauche, $nccR_{x,k,t-1}$ pour le parent droit), ce qui correspond au cas 3 ;
- le pivot complémentaire apparaît dans la partie restante ($nccR_{x,k,t-1}$ pour le parent gauche, $pccR_{x,k,t-1}$ pour le parent droit), ce qui correspond au cas 2.

Les contraintes suivantes imposent cette condition d'admissibilité pour les parents de compensation.

$$\forall t \in \mathcal{S} \setminus \{1\}, \forall x \in \mathcal{X}, \forall k \in \mathcal{C}^t :$$
$$(lcc_{t,k} \wedge v_{x,t}) \implies$$
$$\left(pccF_{x,k,t-1} \vee pccR_{x,k,t-1} \vee nccR_{x,k,t-1}\right) \quad (20)$$
$$(rcc_{t,k} \wedge v_{x,t}) \implies$$
$$\left(nccF_{x,k,t-1} \vee pccR_{x,k,t-1} \vee nccR_{x,k,t-1}\right) \quad (21)$$

Les contraintes (20) et (21) garantissent donc qu'un ensemble de compensation sélectionné est admissible comme parent pour le pivot choisi. La clause concrète extraite de l'ensemble de compensation, ainsi que la mise à jour de sa représentation (F', R'), sont ensuite déterminées par les contraintes de transition introduites ci-dessous.
De plus, la variable pivot elle-même est éliminée par résolution. Elle ne doit donc pas apparaître dans la résolvante obtenue, ni dans la partie restante des ensembles de compensation générés, comme précisé dans les contraintes (22) à (24).

$$\forall t \in \mathcal{S}, \forall x \in \mathcal{X} :$$

$$v_{x,t} \implies \overline{ps_{x,M+t}} \wedge \overline{ns_{x,M+t}} \tag{22}$$

$$v_{x,t} \implies \overline{pccR_{x,2t-1,t}} \wedge \overline{nccR_{x,2t-1,t}} \tag{23}$$

$$v_{x,t} \implies \overline{pccR_{x,2t,t}} \wedge \overline{nccR_{x,2t,t}} \tag{24}$$

Nous nous focalisons maintenant sur la sémantique de transition après l'application d'une étape de max-résolution. Nous distinguons ici deux cas : (i) une clause standard comme parent et (ii) un ensemble de compensation comme parent.

(i) Clause standard comme parent. Si une clause standard est sélectionnée comme parent, le littéral pivot est transféré vers la partie fixe de l'ensemble de compensation correspondant, tandis que chaque littéral non pivot est copié à la fois dans la résolvante et dans la partie restante de cet ensemble de compensation. Les contraintes suivantes décrivent cette mise à jour pour une clause standard sélectionnée comme parent gauche.

$$\forall t \in \mathcal{S}, \forall i \in \mathcal{D}^t, \forall x \in \mathcal{X} :$$

$$(ls_{t,i} \wedge ps_{x,i} \wedge v_{x,t}) \implies pccF_{x,2t-1,t} \tag{25}$$

$$(ls_{t,i} \wedge ps_{x,i} \wedge \overline{v_{x,t}}) \implies (ps_{x,M+t} \wedge pccR_{x,2t-1,t}) \tag{26}$$

$$(ls_{t,i} \wedge ns_{x,i} \wedge \overline{v_{x,t}}) \implies (ns_{x,M+t} \wedge nccR_{x,2t-1,t}) \tag{27}$$

$$(ls_{t,i} \wedge \overline{ps_{x,i}} \wedge \overline{ns_{x,i}}) \implies (\overline{pccF_{x,2t-1,t}} \wedge \overline{nccF_{x,2t-1,t}} \wedge \overline{pccR_{x,2t-1,t}} \wedge \overline{nccR_{x,2t-1,t}}) \tag{28}$$

La contrainte (25) stocke le littéral pivot dans la partie fixe de l'ensemble de compensation gauche généré à l'étape t. Les contraintes (26) et (27) copient chaque littéral non pivot du parent gauche sélectionné à la fois dans la résolvante et dans la partie restante de l'ensemble de compensation gauche généré. Enfin, la contrainte (28) garantit que les variables absentes du parent sélectionné n'apparaissent pas dans l'ensemble de compensation gauche généré.
Les contraintes pour une clause standard sélectionnée comme parent droit sont obtenues de manière symétrique, en stockant le pivot négatif dans la partie fixe de l'ensemble de compensation droit $2t$, et en copiant tous les littéraux non pivots dans la résolvante et dans la partie restante correspondante.

(ii) Ensemble de compensation comme parent. Lorsqu'un ensemble de compensation k est sélectionné comme parent, une variable doit être consommée depuis sa partie restante. Avec la contrainte (6), introduite dans la sémantique de sélection, les contraintes suivantes imposent qu'exactement une telle variable soit consommée. De plus, une variable ne peut être consommée que si au moins l'une de ses polarités apparaît dans la partie restante de l'ensemble de compensation sélectionné.

$$\forall t \in \mathcal{S}, \forall k \in \mathcal{C}^t :$$

$$lcc_{t,k} \implies \bigvee_{x \in \mathcal{X}} accx_{x,t,k} \tag{29}$$

$$rcc_{t,k} \implies \bigvee_{x \in \mathcal{X}} accx_{x,t,k} \tag{30}$$

$$\forall t \in \mathcal{S}, \forall k \in \mathcal{C}^t, \forall x \in \mathcal{X} :$$

$$accx_{x,t,k} \implies (pccR_{x,k,t-1} \vee nccR_{x,k,t-1}) \tag{31}$$

Les contraintes (29) et (30) exigent qu'un parent de compensation sélectionné consomme une variable depuis sa partie restante. Combinées avec la contrainte (6), elles garantissent qu'exactement une variable est consommée. La contrainte (31) garantit que cette consommation est bien définie, c'est-à-dire que la variable consommée apparaît réellement dans la partie restante de l'ensemble de compensation sélectionné.
Si le pivot complémentaire est directement disponible dans la partie restante, il doit être consommé. En revanche, si le pivot apparaît dans la partie restante avec la polarité requise par le côté sélectionné, alors la consommation directe du pivot lui-même est interdite. Dans ce cas, la transition de plus grande clause doit être utilisée, et un autre littéral de la partie restante doit être consommé.

$$\forall t \in \mathcal{S}, \forall k \in \mathcal{C}^t, \forall x \in \mathcal{X} :$$

$$(lcc_{t,k} \wedge v_{x,t} \wedge nccR_{x,k,t-1}) \implies accx_{x,t,k} \tag{32}$$

$$(rcc_{t,k} \wedge v_{x,t} \wedge pccR_{x,k,t-1}) \implies accx_{x,t,k} \tag{33}$$

$$(lcc_{t,k} \wedge v_{x,t} \wedge pccR_{x,k,t-1}) \implies \overline{accx_{x,t,k}} \tag{34}$$

$$(rcc_{t,k} \wedge v_{x,t} \wedge nccR_{x,k,t-1}) \implies \overline{accx_{x,t,k}} \tag{35}$$

$$(lcc_{t,k} \wedge \overline{pccR_{x,k,t-1}} \wedge \overline{nccR_{x,k,t-1}}) \implies \overline{accx_{x,t,k}} \tag{36}$$

$$(rcc_{t,k} \wedge \overline{pccR_{x,k,t-1}} \wedge \overline{nccR_{x,k,t-1}}) \implies \overline{accx_{x,t,k}} \tag{37}$$

$$(\overline{lcc_{t,k}} \wedge \overline{rcc_{t,k}}) \implies \overline{accx_{x,t,k}} \tag{38}$$

Les contraintes (32) et (33) encodent le cas de consommation directe. Du côté gauche, le parent sélectionné doit fournir le pivot positif x. Ainsi, si $\neg x$ apparaît dans la partie restante, il s'agit du pivot complémentaire et il doit être consommé. Symétriquement, du côté droit, le parent sélectionné doit fournir le pivot négatif $\neg x$; donc, si x apparaît dans la partie restante, il doit être consommé. Les contraintes (34) et (35) interdisent la consommation directe du pivot lorsqu'il apparaît dans la partie restante avec la polarité requise par le côté sélectionné. Cela correspond à la transition de plus grande clause : le pivot reste disponible

dans la clause extraite, et un autre littéral de la partie restante doit être consommé à sa place. Les contraintes (36) et (37) indiquent qu'aucun littéral sur la variable x ne peut être consommé depuis un parent de compensation sélectionné si aucune polarité de x n'apparaît dans sa partie restante. Enfin, la contrainte (38) garantit qu'aucune consommation n'est associée à un objet de compensation qui n'est pas sélectionné comme parent à l'étape t.

Les contraintes suivantes, c'est-à-dire (39) à (42), décrivent comment la partie fixe d'un parent de compensation gauche sélectionné est propagée. Si un littéral appartient déjà à la partie fixe de l'ensemble de compensation sélectionné, alors il reste fixé dans le nouvel ensemble de compensation gauche généré. De plus, si ce littéral ne porte pas sur la variable pivot, il est aussi copié dans la résolvante. Cela reflète le fait que les littéraux fixes sont présents dans toutes les clauses représentées par l'objet de compensation, et contribuent donc à la clause parent extraite, sauf s'ils sont éliminés comme pivot.

$$\forall t \in \mathcal{S}, \forall k \in \mathcal{C}^t, \forall x \in \mathcal{X} :$$

$$(lcc_{t,k} \wedge pccF_{x,k,t-1}) \implies pccF_{x,2t-1,t} \quad (39)$$

$$(lcc_{t,k} \wedge nccF_{x,k,t-1}) \implies nccF_{x,2t-1,t} \quad (40)$$

$$(lcc_{t,k} \wedge pccF_{x,k,t-1} \wedge \overline{v_{x,t}}) \implies ps_{x,M+t} \quad (41)$$

$$(lcc_{t,k} \wedge nccF_{x,k,t-1} \wedge \overline{v_{x,t}}) \implies ns_{x,M+t} \quad (42)$$

Les contraintes suivantes, (43) à (46), précisent l'effet de la consommation d'un littéral depuis la partie restante d'un parent de compensation gauche sélectionné. Lorsqu'un littéral positif sur la variable x est consommé depuis la partie restante, le littéral négatif correspondant est ajouté à la partie fixe du nouvel ensemble de compensation. Symétriquement, consommer un littéral négatif ajoute le littéral positif à la partie fixe. Si la variable consommée n'est pas le pivot, le littéral complémentaire est aussi ajouté à la résolvante. Ainsi, ces contraintes encodent la mise à jour irréversible de l'objet de compensation après la consommation d'un littéral de sa partie restante.

$$\forall t \in \mathcal{S}, \forall k \in \mathcal{C}^t, \forall x \in \mathcal{X} :$$

$$(accx_{x,t,k} \wedge pccR_{x,k,t-1}) \implies nccF_{x,2t-1,t} \quad (43)$$

$$(accx_{x,t,k} \wedge nccR_{x,k,t-1}) \implies pccF_{x,2t-1,t} \quad (44)$$

$$(accx_{x,t,k} \wedge pccR_{x,k,t-1} \wedge \overline{v_{x,t}}) \implies ns_{x,M+t} \quad (45)$$

$$(accx_{x,t,k} \wedge nccR_{x,k,t-1} \wedge \overline{v_{x,t}}) \implies ps_{x,M+t} \quad (46)$$

Les contraintes (47) et (48) garantissent que chaque littéral de la partie restante qui n'est pas consommé persiste dans la partie restante du nouvel ensemble de compensation gauche généré. Ainsi, la transition retire uniquement la variable consommée de la partie restante active, tout en conservant les autres littéraux disponibles pour les étapes d'inférence futures.

$$\forall t \in \mathcal{S}, \forall k \in \mathcal{C}^t, \forall x \in \mathcal{X} :$$

$$(lcc_{t,k} \wedge \overline{accx_{x,t,k}} \wedge pccR_{x,k,t-1}) \implies pccR_{x,2t-1,t} \quad (47)$$

$$(lcc_{t,k} \wedge \overline{accx_{x,t,k}} \wedge nccR_{x,k,t-1}) \implies nccR_{x,2t-1,t} \quad (48)$$

Les contraintes (49) et (50) encodent la transition de plus grande clause pour un parent de compensation gauche sélectionné. Lorsque le pivot est justifié par son occurrence dans la partie restante avec la polarité requise, la variable consommée n'est pas le pivot lui-même. Ainsi, tout autre littéral non consommé de la partie restante est copié dans la résolvante. Cela réalise l'extraction de la plus grande clause décrite dans la section 3, tout en évitant la création de plusieurs objets de compensation successeurs.

$$\forall t \in \mathcal{S}, \forall k \in \mathcal{C}^t, \forall x, y \in \mathcal{X},\ x \neq y :$$

$$(lcc_{t,k} \wedge \overline{accx_{x,t,k}} \wedge pccR_{x,k,t-1} \wedge v_{y,t} \wedge pccR_{y,k,t-1}) \implies ps_{x,M+t} \quad (49)$$

$$(lcc_{t,k} \wedge \overline{accx_{x,t,k}} \wedge nccR_{x,k,t-1} \wedge v_{y,t} \wedge pccR_{y,k,t-1}) \implies ns_{x,M+t} \quad (50)$$

Enfin, la contrainte (51) traite les variables qui n'apparaissent pas dans le parent de compensation sélectionné. Si aucune polarité d'une variable n'apparaît dans la partie fixe ni dans la partie restante de l'ensemble de compensation parent, alors aucune polarité n'est introduite dans le nouvel ensemble de compensation gauche généré. Cela empêche la création de littéraux parasites pendant la transition.

$$\forall t \in \mathcal{S}, \forall k \in \mathcal{C}^t, \forall x \in \mathcal{X} :$$

$$(lcc_{t,k} \wedge \neg pccF_{x,k,t-1} \wedge \neg nccF_{x,k,t-1} \wedge \neg pccR_{x,k,t-1} \wedge \neg nccR_{x,k,t-1}) \implies (\neg pccF_{x,2t-1,t} \wedge \neg nccF_{x,2t-1,t} \wedge \neg pccR_{x,2t-1,t} \wedge \neg nccR_{x,2t-1,t}) \quad (51)$$

Notons que les règles de transition omises pour un parent de compensation droit sont obtenues par symétrie avec le cas gauche.

4.5 Persistance conditionnelle et dynamique destructive

Si un ensemble de compensation n'est pas sélectionné comme parent à l'étape t, son contenu persiste à l'étape $t+1$. Les contraintes suivantes implémentent cette condition séparément pour la partie fixe et la partie restante, à la fois pour les littéraux présents et pour les variables absentes.

$$\forall t \in \{1, \ldots, K-1\}, \forall k \in \mathcal{C}, \forall x \in \mathcal{X} :$$

$$(\neg lcc_{t,k} \wedge \neg rcc_{t,k} \wedge pccF_{x,k,t}) \implies pccF_{x,k,t+1} \quad (52)$$

$$(\neg lcc_{t,k} \wedge \neg rcc_{t,k} \wedge nccF_{x,k,t}) \implies nccF_{x,k,t+1} \quad (53)$$

$$(\neg lcc_{t,k} \wedge \neg rcc_{t,k} \wedge pccR_{x,k,t}) \implies pccR_{x,k,t+1} \quad (54)$$

$$(\neg lcc_{t,k} \wedge \neg rcc_{t,k} \wedge nccR_{x,k,t}) \implies nccR_{x,k,t+1} \quad (55)$$

$$(\neg lcc_{t,k} \wedge \neg rcc_{t,k} \wedge \neg pccF_{x,k,t} \wedge \neg nccF_{x,k,t}) \implies (\neg pccF_{x,k,t+1} \wedge \neg nccF_{x,k,t+1}) \quad (56)$$

$$(\neg pccR_{x,k,t} \wedge \neg nccR_{x,k,t}) \implies (\neg pccR_{x,k,t+1} \wedge \neg nccR_{x,k,t+1}) \quad (57)$$

Les contraintes (52) à (55) encodent la persistance positive des ensembles de compensation non sélectionnés. Si un ensemble de compensation k n'est utilisé ni comme parent gauche ni comme parent droit à l'étape t, alors chaque littéral présent dans sa partie fixe ou dans sa partie restante est

préservé à l'étape $t+1$. Ainsi, les objets de compensation non utilisés restent inchangés au fil des étapes d'inférence. Les contraintes (56) et (57) encodent la persistance négative correspondante. Si aucune polarité d'une variable x n'apparaît dans la partie fixe, respectivement dans la partie restante, à l'étape t, alors aucune telle occurrence n'est introduite à l'étape $t+1$ uniquement par persistance. Ensemble, ces contraintes garantissent que le contenu d'un ensemble de compensation ne peut changer que lorsque cet ensemble est sélectionné comme parent ; sinon, la présence et l'absence des littéraux sont propagées à l'étape suivante.
De plus, l'encodage impose la dynamique destructive requise par la sémantique de la max-résolution. En particulier, une clause standard est consommée dès qu'elle est utilisée comme parent, comme décrit par les contraintes (58) et (59).

$$\forall t \in \{1,\dots,K-1\}, \forall i \in \mathcal{D}^t :$$
$$(ls_{t,i} \vee rs_{t,i}) \implies \overline{as_{t+1,i}} \tag{58}$$
$$\overline{as_{t,i}} \implies \overline{as_{t+1,i}} \tag{59}$$

Pour les ensembles de compensation, la disponibilité dépend de leur partie restante. Plus précisément, un ensemble de compensation reste disponible tant que sa partie restante est non vide, comme indiqué dans les contraintes (60) et (61).

$$\forall t \in \{1,\dots,K-1\}, \forall k \in \mathcal{C} :$$
$$\left(\bigwedge_{x\in\mathcal{X}} \overline{pccR_{x,k,t}} \wedge \overline{nccR_{x,k,t}}\right) \implies \overline{acc_{t+1,k}} \tag{60}$$
$$\overline{acc_{t,k}} \implies \overline{acc_{t+1,k}} \tag{61}$$

Ainsi, les contraintes (52) à (61) imposent une évolution temporelle cohérente des objets de preuve. Les ensembles de compensation non sélectionnés persistent sans changement. Les clauses standards sélectionnées sont consommées. Les ensembles de compensation restent disponibles uniquement tant qu'ils contiennent encore des informations actives dans leur partie restante.

4.5.1 Condition de réfutation

Rappelons que nous traitons ici des max-réfutations. Nous devons donc garantir que la dernière clause déduite est vide. La contrainte (62) assure cette propriété en vérifiant que la dernière résolvante ne contient aucun littéral.

$$\overline{ps_{x,M+K}} \wedge \overline{ns_{x,M+K}} \qquad \forall x \in \mathcal{X} \tag{62}$$

5 Expérimentations

Nous menons des tests préliminaires afin d'évaluer la validité et l'efficacité de notre cadre pour identifier les plus courtes max-réfutations. La méthode a été implémenté en Python en utilisant la bibliothèque PySAT [15]. Nous avons encodé les contraintes de cardinalité en FNC à l'aide de réseaux de cardinalité [1]. Pour chaque valeur de K, la formule FNC correspondante Φ_K a été générée puis résolue avec le solveur SAT Cadical [3]. La recherche a été effectuée de manière incrémentale : pour une formule d'entrée donnée, nous résolvons $\Phi_1, \Phi_2, \dots$ jusqu'à trouver le premier encodage satisfiable. Cette première valeur satisfiable est notée $K^\star$. Lorsqu'une affectation satisfaisante est retournée, nous extrayons la preuve correspondante et nous la validons avec un vérificateur indépendant. Ce vérificateur vérifie les parents sélectionnés, la cohérence du pivot, la correction de chaque résolvante et la dérivation finale de la clause vide. Toutes les expérimentations ont été réalisées sur une machine équipée d'un processeur Intel Core i5 à 2.4GHz, avec 4 cœurs et 8GB de RAM.

TABLE 1 – Recherche incrémentale de la plus courte réfutation de ϕ dans l'exemple 1. La longueur minimale est $K^\star = 3$. La ligne $K = 4$ est reportée comme une exécution supplémentaire de validation.

Étape (K)	Statut	Variables CNF	Clauses CNF
$K = 1$	UNSAT	93	343
$K = 2$	UNSAT	177	734
$K^\star = 3$	SAT	276	1205
$K = 4$	SAT	390	1756

Nous avons d'abord considéré la formule $\phi = \{(\neg x_1 \vee x_3), (x_1), (\neg x_1 \vee x_2), (\neg x_2 \vee \neg x_3)\}$ de l'exemple 1. Les résultats du solveur au fil des itérations sont résumés dans le tableau 1. Le modèle a correctement identifié $K = 3$ comme la plus courte longueur de réfutation, avec une preuve plus courte que celles montrées dans la figure 1, tirées de [8]. Cela confirme la correction de notre sémantique d'inférence. Nous avons extrait la preuve à partir de l'affectation retournée par le solveur lors de la dernière itération afin d'obtenir la max-réfutation illustrée dans la figure 2.

Nous avons également effectué une exécution supplémentaire de validation avec $K = 4$. Cette exécution ne fait pas partie de la recherche, puisque le minimum est déjà atteint pour $K^* = 3$. Elle sert plutôt à vérifier que l'encodage peut retrouver l'une des dérivations montrées dans la figure 1, qui utilise une inférence sur une clause de compensation. Dans cette exécution, le solveur retourne SAT et produit une dérivation dans laquelle la clause de compensation $(x_1 \vee \neg x_3)$, générée en résolvant (x_1) avec $(\neg x_1 \vee x_3)$, est ensuite sélectionnée comme prémisse pour dériver $(x_2 \vee \neg x_3)$, c'est à dire qu'on obtien une preuve identique à celle illustrée à gauche dans la figure 1. Cette expérience supplémentaire confirme que note encodage permet de capturer des max-réfutations qui utilisent explicitement des inférences basées sur les clauses de compensation.

Ensuite, nous avons considéré des instances issues du principe des tiroirs. Rappelons que ce problème affirme que n pigeons doivent être placés dans $1 \geq h < n$ tiroirs sans qu'au moins deux pigeons partagent le même tiroir. Nous notons $PHP(m,h)$ la formule indiquant que m pigeons doivent être placés dans h tiroirs, chaque pigeon étant affecté à au moins un tiroir, et aucun deux pigeons ne partageant le même tiroir. Cette famille est un benchmark classique en complexité des preuves, et elle est connue pour être difficile pour les systèmes basés sur la résolution, y compris pour la max-résolution [6].

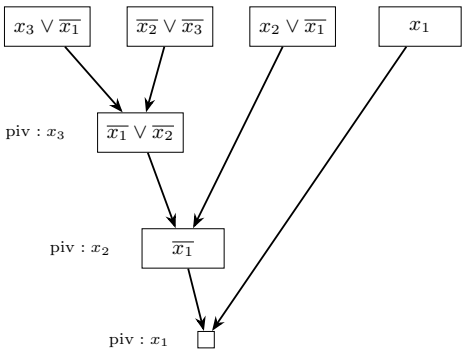


FIGURE 2 – Plus courte max-réfutation pour la formule de l'exemple 1 (les clauses de compensation sont omises, car elles ne sont pas utilisées).

TABLE 2 – Résultats expérimentaux pour l'encodage restreint vérifié de la max-résolution.

Instance	$\|V\|$	$\|C\|$	**Lit.**	$K^\star$	**Temps (s)**	**Valide**
Exemple 1	3	4	7	3	0.0	✓
$PHP(2,1)$	2	3	4	2	0.0	✓
$PHP(3,1)$	3	6	9	2	0.0	✓
$PHP(4,1)$	4	10	16	2	0.1	✓
$PHP(5,1)$	5	15	25	2	0.1	✓
$PHP(3,2)$	6	9	18	10	1800.0	✓
$PHP(4,2)$	8	16	32	10	3635.4	✓

Pour la plus petite instance du principe des tiroirs, $PHP(2,1)$, le cadre identifie correctement une plus courte réfutation de taille $K^\star = 2$. En effet, la formule contient les clauses (x_1), (x_2) et $(\neg x_1 \vee \neg x_2)$; une première étape de résolution dérive une clause unitaire négative, puis une deuxième étape dérive la clause vide. La même longueur optimale est obtenue pour $PHP(3,1)$, car deux pigeons suffisent déjà pour obtenir une contradiction lorsqu'il n'y a qu'un seul tiroir. L'instance $PHP(3,2)$ est beaucoup plus difficile. Le solveur prouve qu'aucune réfutation n'existe pour $K < 10$ dans l'espace de recherche encodé, puis trouve une réfutation correcte pour $K = 10$. À partir des clauses initiales du principe des tiroirs, le certificat extrait dérive la clause vide à travers une séquence de dix étapes de résolution. Pour $PHP(4,2)$, le cadre trouve également une réfutation de longueur $K = 10$. Bien que cette instance contienne plus de variables et de clauses, le certificat extrait utilise seulement un sous-ensemble des clauses initiales, en particulier celles liées au nombre de tiroirs h, correspondant à un cœur insatisfiable déjà présent dans la formule. Cela explique pourquoi $PHP(3,2)$ et $PHP(4,2)$ admettent toutes deux des réfutations de même longueur dans le cadre actuel fondé sur le cœur de résolution.

Les résultats complets sont résumés dans le tableau 2. Pour chaque instance, nous indiquons le nombre de variables $|V|$, le nombre de clauses $|C|$, le nombre total de littéraux, la longueur de la plus courte réfutation $K^\star$, le temps total d'exécution, ainsi que la validation ou non de la preuve extraite par le vérificateur indépendant. Dans l'ensemble, ces expériences fournissent une première preuve de concept pour le cadre proposé. Le solveur réussit à prouver des bornes inférieures pour toutes les valeurs $K < K^\star$, trouve une réfutation pour $K^\star$, et produit des preuves acceptées par un vérificateur indépendant.

6 Conclusion

Dans cet article, nous avons introduit un cadre formel basé sur SAT pour identifier les plus courtes réfutations par max-résolution. Nos résultats préliminaires confirment que le modèle est correct et qu'il est capable de calculer des max-réfutations de taille minimale pour des petites formules. Plusieurs pistes restent à explorer dans les travaux futurs. Tout d'abord, nous prévoyons d'étendre notre évaluation expérimentale en considérant un ensemble plus large d'instances, des encodages de cardinalité plus compacts et des solveurs plus compétitifs. De plus, nous souhaitons généraliser l'encodage en supprimant notre hypothèse restrictive de plus grande clause, en intégrant des règles simples supplémentaires connues pour améliorer l'efficacité de la max-résolution [19], et en étendant notre cadre au calcul des plus courtes preuves d'optimalité pour les instances MaxSAT.

Remerciements

Ce travail est partiellement soutenu par la chaire SAIPS, financée par l'université de Picardie Jules Verne, et par le projet ANR-24-CE23-6126 (BforSAT), financé par l'Agence Nationale de la Recherche.

Références

[1] Roberto Asín, Robert Nieuwenhuis, Albert Oliveras, and Enric Rodríguez-Carbonell. Cardinality networks and their applications. In *International Conference on Theory and Applications of Satisfiability Testing*, pages 167–180. Springer, 2009.

[2] Jeremias Berg, Bart Bogaerts, Jakob Nordström, Andy Oertel, Tobias Paxian, and Dieter Vandesande. Certifying without loss of generality reasoning in solution-improving maximum satisfiability. In Paul Shaw, editor, *30th International Conference on Principles and Practice of Constraint Programming, CP 2024, Girona, Spain, September 2-6, 2024*, volume 307 of *LIPIcs*, pages 4 :1–4 :28. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024.

[3] Armin Biere, Tobias Faller, Katalin Fazekas, Mathias Fleury, Nils Froleyks, and Florian Pollitt. Cadical 2.0. In *International Conference on Computer Aided Verification*, pages 133–152. Springer, 2024.

[4] Armin Biere, Marijn Heule, Hans van Maaren, and Toby Walsh, editors. *Handbook of Satisfiability - Second Edition*, volume 336 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2021.

[5] Maria Luisa Bonet and Jordi Levy. Equivalence between systems stronger than resolution. In Luca Pulina and Martina Seidl, editors, *Theory and Applications of Satisfiability Testing - SAT 2020 - 23rd International Conference, Alghero, Italy, July 3-10, 2020, Proceedings*, volume 12178 of *LNCS*, pages 166–181. Springer, 2020.

[6] Maria Luisa Bonet, Jordi Levy, and Felip Manyà. Resolution for max-sat. *Artificial Intelligence*, 171(8-9) :606–618, 2007.

[7] Mohamed Sami Cherif, Djamal Habet, and André Abramé. Understanding the power of max-sat resolution through up-resilience. *Artif. Intell.*, 289 :103397, 2020.

[8] Mohamed Sami Cherif, Djamal Habet, and Matthieu Py. From crossing-free resolution to max-sat resolution. In Christine Solnon, editor, *28th International Conference on Principles and Practice of Constraint Programming, CP 2022, Haifa, Israel, July 31 - August 8, 2022*, volume 235 of *LIPIcs*, pages 12 :1–12 :17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[9] Sami Cherif, Heythem Sattoutah, Chu-Min Li, Corinne Lucet, and Laure Brisoux-Devendeville. Minimizing Working-Group Conflicts in Conference Session Scheduling Through Maximum Satisfiability. In Paul Shaw, editor, *30th International Conference on Principles and Practice of Constraint Programming (CP 2024)*, volume 307 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 34 :1–34 :11, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[10] Stephen A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, STOC '71, page 151–158, New York, NY, USA, 1971. Association for Computing Machinery.

[11] Ana Graça, Inês Lynce, Joao Marques-Silva, and Arlindo L Oliveira. Efficient and accurate haplotype inference by combining parsimony and pedigree information. In *Algebraic and Numeric Biology : 4th International Conference, ANB 2010, Hagenberg, Austria, July 31-August 2, 2010, Revised Selected Papers*, pages 38–56. Springer, 2012.

[12] João Guerra and Inês Lynce. Reasoning over biological networks using maximum satisfiability. In *International conference on principles and practice of constraint programming*, pages 941–956. Springer, 2012.

[13] Aarti Gupta, Malay K Ganai, and Chao Wang. Sat-based verification methods and applications in hardware verification. In *International School on Formal Methods for the Design of Computer, Communication and Software Systems*, pages 108–143. Springer, 2006.

[14] Federico Heras and João Marques-Silva. Read-once resolution for unsatisfiability-based max-sat algorithms. In Toby Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 572–577. IJCAI/AAAI, 2011.

[15] Alexey Ignatiev, António Morgado, and João Marques-Silva. Pysat : A python toolkit for prototyping with SAT oracles. In Olaf Beyersdorff and Christoph M. Wintersteiger, editors, *Theory and Applications of Satisfiability Testing - SAT 2018 - 21st International Conference, Oxford, UK, July 9-12, 2018, Proceedings*, volume 10929 of *LNCS*, pages 428–437. Springer, 2018.

[16] Javier Larrosa and Federico Heras. Resolution in max-sat and its relation to local consistency in weighted csps. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, page 193–198, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[17] Javier Larrosa, Federico Heras, and Simon De Givry. A logical approach to efficient max-sat solving. *Artificial Intelligence*, 172(2-3) :204–233, 2008.

[18] Javier Larrosa and Emma Rollon. Augmenting the power of (partial) maxsat resolution with extension. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 1561–1568. AAAI Press, 2020.

[19] Javier Larrosa and Emma Rollon. *Towards a Better Understanding of (Partial Weighted) MaxSAT Proof Systems*, volume 12178 of *LNCS*, pages 218–232. Springer, 2020.

[20] Chu Min Li, Felip Manyà, Nouredine Ould Mohamedou, and Jordi Planes. Resolution-based lower bounds in maxsat. *Constraints An Int. J.*, 15(4) :456–484, 2010.

[21] Chu Min Li, Felip Manyà, and Jordi Planes. New inference rules for max-sat. *J. Artif. Intell. Res.*, 30 :321–359, 2007.

[22] Shoulin Li, Jordi Coll, Djamal Habet, Chu-Min Li, and Felip Manyà. *A Tableau Calculus for MaxSAT Based on Resolution*, volume 356 of *Frontiers in Artificial Intelligence and Applications*, pages 35–44. IOS Press, 2022.

[23] Carlos Mencía and Joao Marques-Silva. Computing shortest resolution proofs. In *EPIA Conference on Artificial Intelligence*, pages 539–551. Springer, 2019.

[24] Nina Narodytska and Fahiem Bacchus. Maximum satisfiability using core-guided maxsat resolution. In Carla E. Brodley and Peter Stone, editors, *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pages 2717–2723. AAAI Press, 2014.

[25] Matthieu Py, Mohamed Sami Cherif, and Djamal Habet. Towards bridging the gap between SAT and max-sat refutations. In *32nd IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2020, Baltimore, MD, USA, November 9-11, 2020*, pages 137–144. IEEE, 2020.

[26] Matthieu Py, Mohamed Sami Cherif, and Djamal Habet. A proof builder for max-sat. In Chu-Min Li and Felip Manyà, editors, *Theory and Applications of Satisfiability Testing - SAT 2021 - 24th International Conference, Barcelona, Spain, July 5-9, 2021, Proceedings*, volume 12831 of *LNCS*, pages 488–498. Springer, 2021.

[27] Matthieu Py, Mohamed Sami Cherif, and Djamal Habet. Proofs and certificates for max-sat. *J. Artif. Intell. Res.*, 75 :1373–1400, 2022.

[28] J. A. Robinson. A machine-oriented logic based on the resolution principle. *J. ACM*, 12(1) :23–41, January 1965.

[29] Sean Safarpour, Hratch Mangassarian, Andreas Veneris, Mark H Liffiton, and Karem A Sakallah. Improved design debugging using maximum satisfiability. In *Formal Methods in Computer Aided Design (FMCAD'07)*, pages 13–19. IEEE, 2007.

[30] Irene Unceta, Bernat Salbanya, Jordi Coll, Mateu Villaret, and Jordi Nin. Optimizing resource allocation in home care services using maxsat. *Cogn. Syst. Res.*, 88 :101291, 2024.

[31] Dieter Vandesande, Jordi Coll, and Bart Bogaerts. Certified branch-and-bound maxsat solving (extended version). *CoRR*, abs/2511.10273, 2025.

[32] Dieter Vandesande, Wolf De Wulf, and Bart Bogaerts. Qmaxsatpb : A certified maxsat solver. In Georg Gottlob, Daniela Inclezan, and Marco Maratea, editors, *Logic Programming and Nonmonotonic Reasoning - 16th International Conference, LPNMR 2022, Genova, Italy, September 5-9, 2022, Proceedings*, volume 13416 of *LNCS*, pages 429–442. Springer, 2022.

Satisfiabilité pour le problème du décodage par syndrome à poids élevé

Carl Berton, Sami Cherif, Claire Delaplace

Université de Picardie Jules Verne, Laboratoire MIS UR 4290, Amiens, France

{carl.berton;sami.cherif;claire.delaplace}@u-picardie.fr

Résumé

Le problème du décodage par syndrome à poids élevé (LWSD, pour Large Weight Syndrome Decoding) est un problème fondamental en théorie des codes. Il consiste à déterminer si un code linéaire donné admet un vecteur de poids de Hamming élevé associé à un syndrome spécifique. Le LWSD est une variante du problème classique de décodage par syndrome qui, à l'inverse, recherche une solution de petit poids de Hamming pour un système linéaire défini sur le corps binaire $\mathbb{F}_2$. Dans cet article, nous étudions une généralisation de ce problème au cas d'un corps fini premier $\mathbb{F}_Z$, désignée sous le nom de LWZSD. Nous proposons plusieurs modélisations utilisant des formules de satisfiabilité booléenne (SAT) et comparons l'efficacité de nos approches en utilisant des solveurs de l'état de l'art.

Mots-clés

Décodage par syndrome à poids élevé, Satisfiabilité, Cryptographie

Abstract

The Large Weight Syndrome Decoding problem LWSD is a fundamental problem in coding theory. It consists in determining whether a given linear code admits a high Hamming weight vector associated with a specific syndrome. LWSD is a variant of the classical syndrome decoding problem, which conversely seeks a low Hamming weight solution for a linear system defined over the binary field $\mathbb{F}_2$. In this paper, we investigate a generalization of this problem to the case of a prime finite field $\mathbb{F}_Z$, referred to as LWZSD. We propose several models using Boolean Satisfiability (SAT) formulas and compare the efficiency of our approaches using state-of-the-art solvers.

Keywords

Large Weight Syndrome Decoding, Satisfiability, Cryptography

1 Introduction

La sécurité de l'information est devenue une préoccupation stratégique majeure pour les états, les entreprises et les citoyens, dans un contexte où la multiplication des échanges numériques rend l'usage de la cryptographie essentiel pour garantir la confidentialité, l'authenticité et l'intégrité des données sensibles. Dans ce cadre, la cryptographie symétrique permet une communication confidentielle entre au moins deux parties partageant une clé secrète commune, en utilisant un algorithme de chiffrement symétrique (tel que AES [13]) pour protéger les communications et permettre l'échange sécurisé de clés ou la signature électronique. Cependant, l'émergence des ordinateurs quantiques remet en question la robustesse des schémas classiques, conduisant l'Institut National des Standards et de la Technologie américain (NIST) à lancer un processus de standardisation pour des algorithmes post-quantiques basés sur des schémas asymétriques. Ces derniers s'appuient sur des paires de clés et des problèmes difficiles, tels que le décodage fondé sur les codes, pour assurer une sécurité robuste face aux adversaires tant classiques que quantiques [10].

Parallèlement, les outils de raisonnement automatisé sont devenus de puissants instruments d'analyse et d'optimisation en cryptographie, particulièrement pour affiner les attaques contre les schémas symétriques classiques [12, 21]. Pourtant, de nombreux algorithmes restent peu étudiés sous cet angle et l'application de ces outils aux schémas asymétriques, et particulièrement à ceux résistants aux attaques quantiques, demeure limitée malgré l'existence de problèmes qui se prêtent naturellement à de telles approches. À cet égard, le problème de satisfaisabilité propositionnelle (SAT) s'impose comme un outil de raisonnement automatisé performant, capable d'aider à l'analyse et à l'exploration de constructions cryptographiques post-quantiques. Le problème SAT consiste à déterminer, pour une formule donnée en Forme Normale Conjonctive (CNF), s'il existe une affectation des variables qui la satisfait [8]. Il constitue un problème de décision fondamental en informatique et en intelligence artificielle, utilisé pour résoudre un large éventail de problèmes dans divers domaines tels que la cryptographie [17, 29, 32], la vérification matérielle et logicielle [16], et la planification [24]. Il fut le premier problème démontré comme étant NP-complet [11].

Dans cet article, nous nous intéressons aux problèmes de décodage. Pour assurer une transmission fiable de l'information sur des canaux bruités, les codes correcteurs d'erreurs, et plus spécifiquement les codes linéaires binaires, sont utilisés pour ajouter de la redondance afin de détecter et de corriger les erreurs [26]. Alors que les codes en théorie de l'information sont généralement conçus pour rendre

ce problème facile à résoudre afin de fluidifier les communications, le décodage d'un code linéaire binaire dans le cas général est un problème NP-complet [4]. Les cryptographes s'intéressent à ce problème depuis les années 1970 [19], lorsque McEliece a introduit le premier schéma de chiffrement à clé publique dont la sécurité repose sur la difficulté du décodage d'un code linéaire binaire. Avec l'avènement de la cryptographie post-quantique, la cryptographie fondée sur les codes a connu un essor important ces dernières années, avec de nombreux efforts tant dans les constructions que dans les attaques [20, 15, 22].

Un problème d'un intérêt particulier pour les cryptographes est le décodage par syndrome, qui consiste à retrouver le bruit à partir de la matrice de contrôle de parité du code [3]. En pratique, ce problème revient à résoudre un système linéaire sous-déterminé modulo 2, sous la contrainte que la solution doit avoir un poids de Hamming faible. Des travaux récents ont exploré la modélisation de ce problème à l'aide de formules CNF et XNF (XOR-CNF) [5, 6] et nous poursuivons ces recherches en étudiant le problème du décodage par syndrome à poids élevé (ou LWSD). Contrairement au cas binaire classique, nous considérons ici une généralisation sur un corps fini premier $\mathbb{F}_Z$ (LWZSD). Le passage à $\mathbb{F}_Z$ empêche l'utilisation directe de formules XNF. Nous explorons quatre encodages propositionnels distincts, basés sur des représentations one-hot, unaires et binaires. Ces approches reformulent les contraintes linéaires modulo Z en utilisant des variables auxiliaires et des contraintes Pseudo-Booléennes, tout en intégrant la contrainte de poids de l'erreur. Nous comparons ces différents modèles pour analyser leurs performances respectives avec des solveurs SAT de l'état de l'art.

Cet article est organisé comme suit. Dans la section 2, nous rappelons les concepts fondamentaux nécessaires à notre étude, incluant le problème LWSD et le problème SAT. Nous introduisons nos encodages CNF pour ce problème dans la section 3. Nous analysons ensuite les résultats expérimentaux obtenus pour comparer ces différentes approches dans la section 4. Enfin, nous concluons et discutons des perspectives de travaux futurs dans la section 5.

2 Préliminaires

2.1 Satisfaisabilité

Soit $X = \{x_1, x_2, \ldots, x_n\}$ un ensemble de variables booléennes pouvant prendre les valeurs Vrai ($\top$ ou 1) ou Faux ($\bot$ ou 0). Un littéral est soit une variable $x \in X$, soit sa négation $\overline{x}$. Une formule CNF ϕ est une conjonction de clauses $\phi = C_1 \wedge C_2 \wedge \cdots \wedge C_m$, où chaque clause C_j est une disjonction ($\vee$) de littéraux. Une affectation $\alpha : X \to \{\text{Vrai, Faux}\}$ est une fonction associant à chaque variable de X une valeur de vérité. Ainsi, une clause est satisfaite par une affectation α si au moins l'un de ses littéraux est satisfait par α. Une formule CNF ϕ est satisfaite par une affectation α si toutes ses clauses sont satisfaites par α. Dans ce cas, la formule est dite satisfaisable et α est un modèle de ϕ.

Le problème de satisfaisabilité propositionnelle (SAT) consiste à déterminer si une formule CNF donnée est satisfaisable. Les solveurs SAT modernes sont extrêmement efficaces et parviennent à résoudre des formules comportant un grand nombre de variables et de clauses dans un temps remarquablement court, malgré la difficulté théorique établie du problème. En effet, ces solveurs s'appuient sur l'algorithme CDCL (*Conflict-Driven Clause Learning*) [28] et emploient des mécanismes puissants tels que l'analyse de conflits, le retour arrière non-chronologique (backjumping) et des heuristiques de branchement dédiées [8].

Dans nos modèles, nous utiliserons également des contraintes Pseudo-Booléennes (PB) qui imposent une borne sur une somme pondérée de littéraux $\left(\sum_{i=1}^{h} a_i \times l_i\right) \circ k$ où $a_i, k \in \mathbb{N}$ et $\circ \in \{\leq, =, \geq\}$. Dans le cas où tous les poids sont fixés à 1 (et peuvent donc être omis), de telles contraintes sont plus couramment appelées contraintes de cardinalité. Les contraintes PB et de cardinalité sont typiquement utilisées dans les solveurs pour imposer des bornes pertinentes durant la recherche et peuvent être efficacement encodées sous forme clausale [25].

2.2 Décodage par syndrome à poids élevé

Le problème du décodage par syndrome à poids élevé (LWSD) est une variante fondamentale du problème de décodage par syndrome qui, étant donné un code linéaire binaire, consiste à déterminer s'il existe un mot de code ayant un poids de Hamming élevé. La généralisation du LWSD au cas d'un corps fini premier $\mathbb{F}_Z$ conduit au LWZSD. Nous nous intéressons ainsi à déterminer, pour un code linéaire donné défini sur $\mathbb{F}_Z$, l'existence d'un mot de code à grand poids.

Définition 1 (Poids de Hamming) *Le poids de Hamming* $\mathrm{wt}(\mathbf{e})$ *d'un vecteur* $\mathbf{e} = (e_0, \ldots, e_{n-1}) \in \mathbb{F}_Z^n$ *est défini comme le nombre de ses composantes non nulles, c'est-à-dire :*

$$\mathrm{wt}(\mathbf{e}) := \left|\{\, j \in \{0, \ldots, n-1\} \mid e_j \neq 0 \,\}\right|$$

Définition 2 (LWZSD) *Soit* $\mathbb{F}_Z$ *le corps fini à* Z *éléments, où* Z *est premier, et soient* n, k, *et* t *des entiers tels que* $k \leq n$ *et* $t \leq n$. *Une instance du problème* $\mathsf{LWZSD}(n, k, t)$ *consiste en une matrice de contrôle de parité* $\mathbf{H} \in \mathbb{F}_Z^{(n-k)\times n}$ *et un vecteur* $\mathbf{s} \in \mathbb{F}_Z^{n-k}$ *(appelé le syndrome). Une solution à ce problème est un vecteur* $\mathbf{e} \in \mathbb{F}_Z^n$ *tel que* $\mathrm{He}^\top \equiv \mathrm{s}^\top$ *avec* $wt(e) \geq t$.

Exemple 1 (Décodage par syndrome à poids élevé) *Considérons un code linéaire sur* $\mathbb{F}_3$ *avec les paramètres* $n = 5$, $k = 2$, *et un seuil de poids minimal* $t = 3$. *La matrice de contrôle de parité* $H \in \mathbb{F}_3^{3\times 5}$ *est donnée sous forme systématique* $[I_3 \mid A]$ *comme suit :*

$$H = \begin{pmatrix} 1 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 2 & 1 \\ 0 & 0 & 1 & 1 & 2 \end{pmatrix}$$

Considérons le syndrome $s = (0, 1, 1)^\top \in \mathbb{F}_3^3$. *Nous cherchons un vecteur* $e \in \mathbb{F}_3^5$ *vérifiant* $He^\top \equiv s^\top \pmod 3$ *avec* $\mathrm{wt}(e) \geq t = 3$.

1. ***Proposer un vecteur candidat*** e ***:*** *Contrairement au décodage par syndrome classique qui recherche un vecteur d'erreur de poids faible, il s'agit ici de trouver une solution de poids élevé. Considérons le candidat :*
$$e = (1, 0, 2, 1, 2)^\top$$
qui est de poids $\mathrm{wt}(e) = 4 \geq 3$*, satisfaisant ainsi la contrainte de poids.*
2. ***Vérifier l'équation de syndrome :*** *On calcule* $He^\top \pmod 3$ *:*
$$He^\top = \begin{pmatrix} 6 \\ 4 \\ 7 \end{pmatrix} \equiv \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \pmod 3$$
On obtient bien $He^\top \equiv s^\top \pmod 3$*, ce qui confirme que* e *est une solution valide.*

Le vecteur $e = (1, 0, 2, 1, 2)^\top$ *satisfait les deux conditions du problème* $\mathsf{LWZSD}(5, 2, 3)$ *: il est cohérent avec le syndrome* s *et son poids* $4 \geq t = 3$*.*

Si l'on note $H_{i,j}$ le coefficient à la i-ème ligne et j-ème colonne de la matrice de contrôle H, et s_i la i-ème composante du syndrome, $\mathsf{LWZSD}(n, k, t)$ peut alors être défini comme suit :

$$E_i : \sum_{\substack{0 \leq j \leq n-1 \\ H_{i,j} \neq 0}} H_{i,j} \cdot e_j \equiv s_i \mod Z, \quad \forall 0 \leq i \leq n-k-1 \tag{1a}$$

$$\mathrm{wt}(\mathbf{e}) \geq t \tag{1b}$$

L'équation (1a) exige que, pour chaque E_i, la somme pondérée des composantes du vecteur d'erreur $\mathbf{e}$ soit congrue au syndrome s_i modulo Z, tandis que l'équation (1b) garantit que $\mathbf{e}$ a un poids de Hamming supérieur ou égal à t.
Le schéma de signature Wave [14] illustre l'une des premières constructions cryptographiques reposant explicitement sur la difficulté du décodage par syndrome sur $\mathbb{F}_3$ pour des mots d'erreur à grand poids. Bricout et al. [9] ont étudié ce problème en détail, soulignant que la structure combinatoire du décodage ternaire à grand poids diffère significativement de celle du décodage binaire classique. Leur analyse exploite des algorithmes basés sur les techniques de décodage par ensemble d'information (ISD) [23, 31, 2] et montre que le cas du poids élevé constitue un cadre pertinent pour les applications cryptographiques. En effet, ce problème est considéré comme une alternative crédible face à la menace quantique, car les algorithmes de Shor [27] ne s'appliquent pas à la structure des codes correcteurs d'erreurs.
Plus récemment, le décodage par syndrome classique a été étudié sous l'angle de la satisfaisabilité propositionnelle. Des travaux antérieurs ont abouti à des modèles SAT sous forme de formules CNF et XNF (XOR-CNF) [5, 6]. Cependant, ces travaux se focalisent sur le cas $\mathbb{F}_2$ avec un poids de Hamming borné, ce qui permettait d'exploiter naturellement la structure XOR d'origine du problème pour introduire des modèles ensuite traduits en CNF. À l'inverse, le problème LWZSD implique des corps $\mathbb{F}_Z$ où $Z > 2$, ce qui limite l'usage direct et naturel des contraintes XOR. Par conséquent, les modèles XNF spécifiques au cas binaire ne sont pas directement transposables à notre cadre, nécessitant l'étude d'encodages CNF dédiés, qui seront présentés dans la section suivante.

3 Modélisation du décodage par syndrome à poids élevé

3.1 Variables et notations

Dans cette section, nous introduisons les variables, ensembles et notations qui seront utilisés dans les sections suivantes pour décrire les différents modèles. Nous notons E_i la i-ème équation du système, pour $0 \leq i \leq n-k-1$. Pour simplifier les notations, posons $m = n-k-1$. Pour encoder chaque $e_j \in \{0, \ldots, Z-1\}$, nous introduisons les variables booléennes $e_{j,1}, e_{j,2}, \ldots, e_{j,Z-1}$, pour $0 \leq j \leq n-1$, où $e_{j,k} = 1 \Leftrightarrow e_j \geq k$. En d'autres termes, si $e_j = v$, alors $e_{j,1} = \cdots = e_{j,v} = 1$ et $e_{j,v+1} = \cdots = e_{j,Z-1} = 0$. Pour chaque équation E_i, nous définissons l'ensemble V_{E_i} comme l'ensemble des indices des variables apparaissant dans E_i. Plus formellement, pour tout $i \in \{0, \ldots, m\}$, nous avons :

$$V_{E_i} = \{j \in \{0, \ldots, n-1\} \mid H_{i,j} \neq 0\}$$

Pour chaque équation E_i, nous considérons la somme pondérée des variables qui y apparaissent, à savoir $\sum_{j \in V_{E_i}} H_{i,j} e_j$. Dans ce contexte, nous introduisons pour chaque équation un ensemble noté K_{E_i}, correspondant à l'ensemble des valeurs entières que cette somme peut prendre tout en satisfaisant la congruence imposée par s_i. Pour caractériser cet ensemble, nous commençons par déterminer une borne inférieure. Puisque le vecteur d'erreur e doit satisfaire $\mathrm{wt}(e) \geq t$, ce qui signifie qu'il contient au moins t composantes non nulles, cela implique qu'au plus $n-t$ composantes de e sont nulles parmi les variables participantes. Par conséquent, nous posons $m_i = \max\big(0,\ |V_{E_i}| - (n-t)\big)$, ce qui représente le nombre minimal de composantes non nulles devant figurer dans E_i. Nous définissons ensuite $V_{E_i}^{\min}$ comme l'ensemble des indices correspondant aux m_i plus petits coefficients $H_{i,j}$ parmi $j \in V_{E_i}$. Ainsi, nous pouvons définir formellement la borne inférieure $v_{\min}^{(i)} \in \mathbb{N}$ par :

$$v_{\min}^{(i)} = \min\Big\{v \geq \sum_{j \in V_{E_i}^{\min}} H_{i,j} \Big|\ v \equiv s_i \pmod Z\Big\}$$

En d'autres termes, $v_{\min}^{(i)}$ est la plus petite valeur entière congrue à s_i modulo Z et supérieure ou égale à la somme minimale réalisable. De même, nous pouvons déterminer une borne supérieure pour l'ensemble K_{E_i}. En effet, chaque variable e_j prend une valeur entière entre 0 et $Z-1$. Ainsi, la somme pondérée $\sum_{j \in V_{E_i}} H_{i,j} e_j$ atteint sa valeur maximale lorsque toutes les variables e_j sont égales à $Z-1$. Pour que cette valeur appartienne à l'ensemble K_{E_i}, elle

doit également respecter la congruence imposée par s_i modulo Z. Nous définissons donc formellement $v_{\max}^{(i)} \in \mathbb{N}$ par :

$$v_{\max}^{(i)} = \max\Big\{v \leq (Z-1) \sum_{j \in V_{E_i}} H_{i,j} \Big|\, v \equiv s_i \pmod{Z}\Big\}$$

En d'autres termes, $v_{\max}^{(i)}$ est la plus grande valeur entière inférieure ou égale à la somme maximale possible et congrue à s_i modulo Z. Nous pouvons alors définir l'ensemble des valeurs possibles que la somme pondérée $\sum_{j \in V_{E_i}} H_{i,j} e_j$ peut prendre dans l'équation E_i, en tenant compte de la parité imposée par s_i dans $\mathbb{F}_Z$ comme suit :

$$K_{E_i} = \Big\{v \in \{\, v_{\min}^{(i)}, \ldots, v_{\max}^{(i)}\} \Big|\, v \equiv s_i \pmod{Z}\Big\}.$$

D'autres variables seront introduites dans les sections suivantes, leur sémantique dépendant du contexte spécifique de chaque modèle.

3.2 Encodage one-hot

Pour représenter la valeur prise par la somme des variables dans chaque équation, nous introduisons des variables $x_{i,v}$ pour tout $(i,v) \in \{0,\ldots,m\} \times K_{E_i}$, qui sont affectées à Vrai si et seulement si la somme des variables de l'équation E_i est égale à v, c'est-à-dire $\sum_{j \in V_{E_i}} H_{i,j}\, e_j = v$. Le système $\mathsf{LWZSD}(n,k,t)$ peut alors s'écrire sous la forme suivante :

$$\sum_{j \in V_{E_i}} \sum_{c=1}^{Z-1} H_{i,j}\, e_{j,c} + \sum_{v \in K_{E_i}} v\, \bar{x}_{i,v} = \sum_{v \in K_{E_i}} v, \quad \forall\, 0 \leq i \leq m \tag{2a}$$

$$\sum_{v \in K_{E_i}} x_{i,v} = 1, \quad \forall\, 0 \leq i \leq m \tag{2b}$$

$$\bar{e}_{j,v} \vee e_{j,v-1}, \quad \forall\, 0 \leq j \leq n-1,\ 2 \leq v \leq Z-1 \tag{2c}$$

$$\sum_{j=0}^{n-1} e_{j,1} \geq t \tag{2d}$$

La contrainte (2a), écrite sous forme de contrainte PB, impose que la somme pondérée des variables d'erreur e_j dans l'équation E_i soit exactement égale à la valeur v choisie par les variables $x_{i,v}$. La contrainte (2b) garantit qu'une seule valeur de somme est sélectionnée pour chaque équation. La contrainte (2c) assure que si $e_{j,v} = 1$, alors tous les $e_{j,c}$ avec $c < v$ sont également égaux à 1. Enfin, la contrainte (2d) exige que le vecteur d'erreur e possède au moins t composantes non nulles.

Proposition 1 *La formule propositionnelle induite par les contraintes* (2) *est satisfaisable si et seulement si le système* $\mathsf{LWZSD}(n,k,t)$ *admet une solution.*

Preuve. Tout d'abord, nous établissons que les contraintes (2a), (2b) et (2c) sont équivalentes à l'équation de syndrome (1a). Soit $i \in \{0,\ldots,m\}$, en notant que $x = 1 - \bar{x}$, la contrainte (2a) peut s'écrire :

$$\sum_{j \in V_{E_i}} H_{i,j} \sum_{c=1}^{Z-1} e_{j,c} + \sum_{v \in K_{E_i}} v\,(1 - x_{i,v}) = \sum_{v \in K_{E_i}} v.$$

Cette égalité est équivalente à :

$$\sum_{j \in V_{E_i}} H_{i,j} \left(\sum_{c=1}^{Z-1} e_{j,c}\right) = \sum_{v \in K_{E_i}} v \cdot x_{i,v}$$

Sous la contrainte (2c), les variables $e_{j,c}$ constituent un encodage de la variable e_j, avec $e_j = \sum_{c=1}^{Z-1} e_{j,c}$. Nous obtenons ainsi :

$$\sum_{j \in V_{E_i}} H_{i,j} \cdot e_j = \sum_{v \in K_{E_i}} v \cdot x_{i,v}.$$

De plus, la contrainte (2b) impose que, pour chaque équation E_i, les variables $x_{i,v}$ forment un encodage one-hot des valeurs de somme possibles modulo Z. Compte tenu de la définition de l'ensemble K_{E_i}, cette contrainte est donc équivalente à l'équation de syndrome (1a) du problème $\mathsf{LWZSD}(n,k,t)$.
Il reste à considérer la contrainte de poids. Par construction, la variable $e_{j,1}$ est vraie si et seulement si $e_j \geq 1$. Ainsi, nous avons :

$$\mathrm{wt}(\mathbf{e}) = \big|\{\, j \in \{0,\ldots,n-1\} \mid e_j \neq 0\,\}\big| = \sum_{j=0}^{n-1} e_{j,1}.$$

La contrainte (2d) est donc équivalente à la contrainte (1b). Ceci établit l'équivalence entre les solutions du système $\mathsf{LWZSD}(n,k,t)$ et les affectations satisfaisant les contraintes (2). □

3.3 Encodage unaire

Nous pouvons également observer que, dans le cas général, les éléments de l'ensemble K_{E_i} augmentent par incréments de Z. Pour tirer parti de cette structure, nous redéfinissons la sémantique des variables $x_{i,v}$ comme suit : $x_{i,v} = 1 \Leftrightarrow \sum_{j \in V_{E_i}} H_{i,j}\, e_j \geq v, \forall v \in K_{E_i}$. Le système peut alors s'écrire ainsi :

$$\sum_{j \in V_{E_i}} \sum_{c=1}^{Z-1} H_{i,j}\, e_{j,c} \;+ \sum_{v \in K_{E_i} \setminus \{v_{\min}^{(i)}\}} Z \cdot \bar{x}_{i,v} = v_{\max}^{(i)}, \quad \forall\, 0 \leq i \leq m \tag{3a}$$

$$\bar{x}_{i,v} \vee x_{i,v-Z}, \quad \forall 0 \leq i \leq m,\ v \in K_{E_i} \setminus \{v_{\min}^{(i)}\} \tag{3b}$$

$$\bar{e}_{j,v} \vee e_{j,v-1}, \quad \forall 0 \leq j \leq n-1,\ 2 \leq v \leq Z-1 \tag{3c}$$

$$\sum_{j=0}^{n-1} e_{j,1} \geq t \tag{3d}$$

La contrainte (3a) exprime, sous forme PB, la relation entre la somme pondérée des variables d'erreur e_j dans l'équation E_i et la valeur sélectionnée parmi les variables $x_{i,v}$. La contrainte (3b) garantit que si $\sum_{j \in V_{E_i}} e_j \geq v$ est vrai,

alors $\sum_{j\in V_{E_i}} e_j \geq v - Z$ est également vrai, assurant ainsi la cohérence de l'encodage. La contrainte (3c) encode la représentation unaire des variables e_j. Elle impose que, dès que $e_{j,v} = 1$, toutes les variables $e_{j,c}$ avec $c < v$ sont également fixées à 1. Enfin, la contrainte de poids (3d) garantit que le vecteur d'erreur e possède au moins t composantes non nulles.

Proposition 2 *La formule propositionnelle induite par les contraintes* (3) *est satisfaisable si et seulement si le système* $\mathsf{LWZSD}(n, k, t)$ *admet une solution.*

Preuve. Nous démontrons que, sous les contraintes (3b) et (3c), la contrainte Pseudo-Booléenne (3a) est équivalente à l'équation de syndrome (1a) du système $\mathsf{LWZSD}(n, k, t)$. En notant que $\overline{x} = 1 - x$, la contrainte (3a) peut être réécrite comme suit :

$$\sum_{j\in V_{E_i}} \sum_{c=1}^{Z-1} H_{i,j}\, e_{j,c} = v_{\max}^{(i)} + \sum_{v\in K_{E_i}\setminus\{v_{\min}^{(i)}\}} Z\cdot(x_{i,v} - 1)$$

On note ici que par définition des ensembles K_{E_i}, nous avons $\sum_{v\in K_{E_i}\setminus\{v_{\min}^{(i)}\}} Z = v_{\max}^{(i)} - v_{\min}^{(i)}$. Ainsi, de manière analogue à la démonstration de la proposition 1, nous pouvons déduire la forme équivalente suivante :

$$\sum_{j\in V_{E_i}} H_{i,j}\cdot e_j = v_{\min}^{(i)} + \sum_{v\in K_{E_i}\setminus\{v_{\min}^{(i)}\}} Z\cdot x_{i,v}$$

Sous la contrainte (3b), la cohérence des variables $x_{i,v}$ pour la sélection des valeurs successives de v dans K_{E_i} par incréments de Z est assurée. Par conséquent, la contrainte (3a) est satisfaite si et seulement si $\sum_{j\in V_{E_i}} H_{i,j}\, e_j$ est égal à l'un des éléments de l'ensemble K_{E_i}, ce qui correspond exactement à la condition imposée par l'équation de syndrome (1a) associée à E_i. □

3.4 Encodage binaire

Dans cette section, notre objectif est de représenter en binaire le quotient de la division euclidienne de la somme $\sum_{j\in V_{E_i}} H_{i,j}e_j$, tout en imposant naturellement la contrainte de congruence fixée par s_i. Pour chaque équation E_i, $0 \leq i \leq m$, cette contrainte implique l'existence d'un unique entier $q \in \mathbb{N}$ tel que :

$$\sum_{j\in V_{E_i}} H_{i,j}\, e_j = s_i + Zq.$$

Nous représentons ensuite q en binaire à l'aide de variables auxiliaires $x_{i,\ell} \in \{0, 1\}$, définies par $x_{i,\ell} = 1$ si et seulement si le ℓ-ème bit de la représentation binaire de q est 1, pour $0 \leq \ell \leq L_i - 1$. Nous notons par $b_i(q) := (x_{i,L_i-1}, x_{i,L_i-2}, \ldots, x_{i,0})$ la représentation binaire de q, où L_i est le nombre minimal de bits nécessaires pour encoder la valeur maximale possible de q dans l'équation E_i. On note $b_{i,\ell}(q)$ le ℓ-ième bit de cette représentation, tel que $b_{i,\ell}(q) = x_{i,\ell}$.

Plus précisément, puisque nous avons défini $v_{\max}^{(i)}$ (resp. $v_{\min}^{(i)}$) comme la somme entière maximale (resp. minimale) possible pour l'équation E_i, nous pouvons introduire les notations suivantes :

$$q_{\max}^{(i)} := \frac{v_{\max}^{(i)} - s_i}{Z} \quad \text{et} \quad q_{\min}^{(i)} := \frac{v_{\min}^{(i)} - s_i}{Z},$$

où $q_{\max}^{(i)}$ (resp. $q_{\min}^{(i)}$) représente le quotient de la division euclidienne de $v_{\max}^{(i)}$ (resp. $v_{\min}^{(i)}$) par Z. Ainsi, nous pouvons représenter le quotient q comme suit, avec L_i représentant le nombre de bits requis pour q :

$$q = \sum_{\ell=0}^{L_i-1} 2^\ell\, x_{i,\ell},\ \text{où}\ L_i = \left\lceil \log_2\left(q_{\max}^{(i)} + 1\right)\right\rceil.$$

Le système $\mathsf{LWZSD}(n, k, t)$ peut donc s'écrire sous la forme suivante :

$$\sum_{j\in V_{E_i}} H_{i,j}e_j + Z\sum_{\ell=0}^{L_i-1} 2^\ell\cdot\bar{x}_{i,\ell} = s_i + Z\sum_{\ell=0}^{L_i-1} 2^\ell,$$

$$\forall\, 0 \leq i \leq m \quad \text{(4a)}$$

$$\bar{e}_{j,v} \vee e_{j,v-1}, \quad \forall 0 \leq j \leq n-1,\ 2 \leq v \leq Z-1 \quad \text{(4b)}$$

$$\sum_{j=0}^{n-1} e_{j,1} \geq t \quad \text{(4c)}$$

La contrainte (4a), écrite sous forme PB, exprime que la somme pondérée des variables d'erreur $\sum_{j\in V_{E_i}} H_{i,j}e_j$ satisfait la relation modulo Z, en utilisant les variables auxiliaires $x_{i,\ell}$ représentant le quotient q en binaire. De nouveau, les contraintes (4b) et (4c) encodent correctement les variables d'erreur et imposent que le vecteur d'erreur possède au moins t composantes non nulles.

Proposition 3 *La formule propositionnelle induite par les contraintes* (4) *est satisfaisable si et seulement si le système* $\mathsf{LWZSD}(n, k, t)$ *admet une solution.*

Preuve. Sous la contrainte (4b), la contrainte Pseudo-Booléenne (4a) est équivalente à l'équation de syndrome (1a) du système $\mathsf{LWZSD}(n, k, t)$. Pour une équation E_i, avec $i \in \{0, \ldots, m\}$, isoler la somme pondérée des variables d'erreur dans (4a) donne :

$$\sum_{j\in V_{E_i}} H_{i,j}\, e_j = s_i + Z\sum_{\ell=0}^{L_i-1} 2^\ell Z + \sum_{\ell=0}^{L_i-1} 2^\ell(x_{i,\ell} - 1)$$

Ainsi, en posant $q := \sum_{\ell=0}^{L_i-1} 2^\ell\, x_{i,\ell}$, nous obtenons la forme suivante :

$$\sum_{j\in V_{E_i}} H_{i,j}\, e_j = s_i + Z\sum_{\ell=0}^{L_i-1} 2^\ell\, x_{i,\ell} = s_i + Zq$$

ce qui correspond à l'équation de syndrome (1a) pour E_i. □

Il convient de noter que l'expression $s_i + Zq$ ne tient pas compte des bornes inférieure et supérieure imposées par le problème initial. En effet, pour chaque équation E_i, seules les valeurs de sommes appartenant à l'ensemble K_{E_i} sont admissibles. Cependant, nous autorisons actuellement également des valeurs qui ne devraient pas l'être, c'est-à-dire des quotients q tels que $s_i + Zq \notin K_{E_i}$. Il est donc nécessaire d'éliminer ces configurations invalides en ajoutant des contraintes complémentaires. Nous présentons ci-dessous deux méthodes pour imposer cette restriction en agissant directement sur le quotient et les variables auxiliaires qui le représentent.

Filtrage exhaustif. Puisque q est représenté à l'aide de L_i bits, la valeur maximale qu'il peut prendre est $2^{L_i}-1$. Selon la construction précédente, les seules valeurs admissibles pour le quotient q sont celles comprises entre $q^{(i)}_{\min}$ et $q^{(i)}_{\max}$. Ainsi, toutes les valeurs situées dans les intervalles suivants doivent être interdites :

$$I^{(i)} = [0, q^{(i)}_{\min}[\ \cup \ ]q^{(i)}_{\max}, 2^{L_i} - 1], \quad \forall\, i \in \{0, \dots, m\}$$

Par conséquent, pour chaque valeur interdite, nous pouvons ajouter une clause qui exclut explicitement l'affectation correspondant à cette valeur. De telles clauses doivent imposer qu'au moins un bit du quotient soit différent et s'écrivent formellement comme suit :

$$\bigvee_{\substack{0\le \ell \le L_i-1 \\ b_{i,\ell}(v)=0}} x_{i,\ell} \ \vee \bigvee_{\substack{0\le \ell \le L_i-1 \\ b_{i,\ell}(v)=1}} \bar{x}_{i,\ell}, \quad \forall i \in \{0,\dots,m\},\, v \in I^{(i)} \tag{5}$$

Exemple 2 *Considérons $v = 12$, dont la représentation binaire sur 4 bits est $b_i(v) = (1,1,0,0)$. La clause résultante pour interdire cette valeur est : $(\bar{x}_{i,3} \vee \bar{x}_{i,2} \vee x_{i,1} \vee x_{i,0})$, ce qui empêche l'affectation $(x_{i,3}, x_{i,2}, x_{i,1}, x_{i,0}) = (1,1,0,0)$.*

Nous obtenons un premier modèle binaire qui combine l'ensemble des contraintes (4) et (5). Pour chaque équation E_i, cette méthode de filtrage exhaustif génère exactement une clause par valeur interdite de q. Afin de réduire le nombre de clauses et d'obtenir un encodage plus compact, nous proposons dans la section suivante une méthode plus judicieuse qui agit directement sur les variables binaires représentant le quotient.

Filtrage compact. Pour commencer, nous définissons l'ensemble $J^{(i)}_{\max}$ (resp. $J^{(i)}_{\min}$) comme l'ensemble des indices où le bit vaut 1 (resp. 0) dans $q^{(i)}_{\max}$ (resp. $q^{(i)}_{\min}$). Dans la suite, nous détaillons uniquement le raisonnement pour la borne supérieure $q^{(i)}_{\max}$. Le traitement de la borne inférieure $q^{(i)}_{\min}$ repose sur un raisonnement symétrique.

Soit $(b_{i,L_i-1}, \dots, b_{i,0})$ la décomposition binaire de $q^{(i)}_{\max}$. Nous identifions l'ensemble des indices de début de bloc, noté $\mathcal{B}^{(i)}_{\max} = \{l \in J^{(i)}_{\max} \mid b_{i,l-1} = 0\}$, qui correspond aux positions minimales de chaque séquence de 1 consécutifs dans la représentation binaire de $q^{(i)}_{\max}$.

Exemple 3 *On considère $q^{(i)}_{\max} = 111001101$. Alors $J^{(i)}_{\max} = \{8,7,6,3,2,0\}$, et l'ensemble $\mathcal{B}^{(i)}_{\max} = \{6,2\}$.*

Nous introduisons aussi des variables auxiliaires $p^{\max}_{i,j}$, qui représentent le préfixe binaire de $q^{(i)}_{\max}$, formé par les bits lus du bit de poids fort vers le bit de poids faible, jusqu'au bit d'indice j inclus. La variable $p^{\max}_{i,j}$ satisfait donc la relation suivante, où j' désigne le premier indice supérieur à j dans $\mathcal{B}^{(i)}_{\max}$:

$$p^{\max}_{i,j} \iff p^{\max}_{i,j'} \wedge \bigwedge_{\substack{j\le l\le j'-1 \\ b_{i,l}=1}} x_{i,l} \wedge \bigwedge_{\substack{j\le k\le j'-1 \\ b_{i,l}=0}} \bar{x}_{i,l}$$

Cette relation est encodée en CNF par les clauses suivantes :

$$\bigwedge_{j\le l\le L_i-1} (\bar{p}^{\max}_{i,j} \vee x_{i,l}),$$

$$\left(p^{\max}_{i,j} \vee \bigvee_{\substack{j\le l\le L_i-1 \\ b_{i,l}=1}} \bar{x}_{i,l}\right), \quad \forall i \in \{0,\dots,m\},$$

$$j = \max(\mathcal{B}^{(i)}_{\max}) \tag{6a}$$

$$(\bar{p}^{\max}_{i,j} \vee p^{\max}_{i,j'}) \wedge \bigwedge_{\substack{j\le l\le j'-1 \\ b_{i,l}=1}} (\bar{p}^{\max}_{i,j} \vee x_{i,l}) \wedge \bigwedge_{\substack{j\le l\le j'-1 \\ b_{i,l}=0}} (\bar{p}^{\max}_{i,j} \vee \bar{x}_{i,l}),$$

$$\left(p^{\max}_{i,j} \vee \bar{p}^{\max}_{i,j'} \vee \bigvee_{\substack{j\le l\le j'-1 \\ b_{i,l}=1}} \bar{x}_{i,l} \vee \bigvee_{\substack{j\le l\le j'-1 \\ b_{i,l}=0}} x_{i,l}\right),$$

$$\forall i \in \{0,\dots,m\},\ \forall j \in \mathcal{B}^{(i)}_{\max} \setminus \{\max(\mathcal{B}^{(i)}_{\max})\},$$

$$j' = \min\{l \in \mathcal{B}^{(i)}_{\max} \mid l > j\} \tag{6b}$$

Rappelons que L_i correspond au nombre minimal de bits requis pour représenter $q^{(i)}_{\max}$ tel que $s_i + Z \cdot q^{(i)}_{\max} = v^{(i)}_{\max}$ soit la plus grande somme entière avant réduction modulo Z appartenant à l'ensemble K_{E_i}. Il est clair que le bit de poids fort $b_{i,L_i-1} = 1$. Ainsi, pour obtenir une valeur strictement supérieure à $q^{(i)}_{\max}$, il est nécessaire de modifier au moins l'un des bits de poids inférieur (indices de 0 à L_i-2). Pour éliminer les valeurs supérieures à $q^{(i)}_{\max}$, nous interdisons d'abord les configurations dans lesquelles, entre deux bits fixés à 1, au moins un bit fixé à 0 est basculé, produisant ainsi une valeur strictement supérieure à $q^{(i)}_{\max}$. Pour empêcher cela, nous pouvons utiliser les clauses suivantes :

$$\bigwedge_{j'+1\le k\le j-1} \left(\bar{p}^{\max}_{i,j} \vee \bar{x}_{i,k}\right), \quad \forall i \in \{0,\dots,m\},$$

$$\forall j \in \mathcal{B}^{(i)}_{\max}, j' = \max\{l \in J^{(i)}_{\max} \mid l < j\} \tag{6c}$$

La contrainte ci-dessus permet au préfixe binaire jusqu'à l'indice j (inclus) d'être identique à celui de $q^{(i)}_{\max}$, tout en

interdisant à tout bit situé strictement entre $j-1$ et $j'+1$ de prendre la valeur 1. Cependant, lorsque nous atteignons le dernier élément de l'ensemble $J^{(i)}_{\max}$, les bits de poids inférieur peuvent encore être égaux à 0. Dans ce cas, il serait possible d'obtenir une valeur strictement supérieure à $q^{(i)}_{\max}$ en remplaçant l'un de ces zéros par un 1. Pour éviter cela, nous générons les clauses suivantes :

$$\bigwedge_{0\le k\le j-1}\left(\bar{p}^{\max}_{i,j}\vee\bar{x}_{i,k}\right),\quad \forall i\in\{0,\dots,m\},$$
$$j=\min\left(J^{(i)}_{\max}\right)\quad \text{si } 0\notin J^{(i)}_{\max}\quad \text{(6d)}$$

En combinant les contraintes (4) et les contraintes clausales du filtrage compact, nous obtenons un modèle qui élimine les valeurs strictement supérieures à $q^{(i)}_{\max}$ (et strictement inférieures à $q^{(i)}_{\min}$ par un raisonnement symétrique). Enfin, notons que pour une équation E_i donnée où L_i représente le nombre de bits nécessaires pour le quotient q, le filtrage exhaustif nécessite $O(2^{L_i})$ clauses sans variables additionnelles, alors que le filtrage compact ne nécessite que $O(L_i)$ clauses et variables.

Exemple 4 *Soit $n=20$, $Z=3$, $t=19$, et considérons l'équation de parité $E_0=[1,0,0,0,0,0,0,0,2,1,2,0,2,2,2,1,1,0,2,1]$ avec $s_0=1$. Les indices des variables apparaissant dans l'équation sont $V_{E_0}=\{0,8,9,10,12,13,14,15,16,18,19\}$, avec $|V_{E_0}|=11$. Comme la contrainte de poids impose $\mathrm{wt}(e)\ge t=19$, il peut y avoir au plus $n-t=1$ composante nulle dans le vecteur d'erreur. Ainsi, nous posons $m_0=\max(0,|V_{E_0}|-(n-t))=10$, ce qui représente le nombre minimal de composantes non nulles devant participer à cette équation.*
Nous pouvons maintenant définir l'ensemble K_{E_0}, qui correspond aux valeurs entières que la somme des variables d'erreur peut prendre tout en satisfaisant la congruence imposée par s_0. Il y a 5 coefficients égaux à 1 et 6 coefficients égaux à 2. Nous définissons alors $V^{\min}_{E_0}=\{0,9,15,16,19,8,10,12,13,14\}$ comme l'ensemble des indices correspondant aux m_0 plus petits coefficients $H_{0,j}$ parmi $j\in V_{E_0}$. Cet ensemble nous permet de déterminer la borne inférieure de l'ensemble K_{E_0}, qui est $\sum_{j\in V^{\min}_{E_0}}H_{0,j}=15$.
Nous savons également que la somme pondérée des variables d'erreur atteint sa valeur maximale lorsque toutes les variables e_j sont égales à $Z-1$; ainsi, la somme entière réalisée par cette équation doit satisfaire $15\le\sum_{j\in V_{E_0}}H_{0,j}e_j\le 34$. Puisque la contrainte modulaire impose $\sum_{j\in V_{E_0}}H_{0,j}e_j\equiv s_0=1\pmod 3$, les valeurs admissibles sont exactement $K_{E_0}=\{v\in[15,34]\mid v\equiv 1\pmod 3\}=\{16,19,22,25,28,31,34\}$. Ainsi, $v^{(0)}_{\min}=16$ et $v^{(0)}_{\max}=34$.
Rappelons que la somme pondérée des variables d'erreur pour une équation E_i s'écrit

$$\sum_{j\in V_{E_i}}H_{i,j}\,e_j=s_i+Z\sum_{\ell=0}^{L_i-1}2^{\ell}x_{i,\ell}\ ,$$

q	$x_{0,3}$	$x_{0,2}$	$x_{0,1}$	$x_{0,0}$	$q\cdot Z+s_0$	État
0	0	0	0	0	1	✗ (6c)
1	0	0	0	1	4	✗ (6c)
2	0	0	1	0	7	✗ (6c)
3	0	0	1	1	10	✗ (6c)
4	0	1	0	0	13	✗ (6d)
5	0	1	0	1	$v^{(0)}_{\min}=16$	✓
6	0	1	1	0	19	✓
7	0	1	1	1	22	✓
8	1	0	0	0	25	✓
9	1	0	0	1	28	✓
10	1	0	1	0	31	✓
11	1	0	1	1	$v^{(0)}_{\max}=34$	✓
12	1	1	0	0	37	✗ (6c)
13	1	1	0	1	40	✗ (6c)
14	1	1	1	0	43	✗ (6c)
15	1	1	1	1	46	✗ (6c)

TABLE 1 – Représentation binaire des quotients et état des sommes $q\cdot Z+s_0$ (✓ : autorisé, ✗ : interdit).

où $q=\sum_{\ell=0}^{L_i-1}2^{\ell}x_{i,\ell}$ représente le quotient, et L_i est choisi pour encoder toutes les valeurs admissibles de q. Nous calculons

$$L_0=\left\lceil\log_2\left(\left\lfloor\frac{v^{(0)}_{\max}-s_0}{Z}\right\rfloor+1\right)\right\rceil=4$$

et, par conséquent, seuls quatre bits sont nécessaires pour représenter toutes les valeurs possibles du quotient. Le Tableau 1 illustre le processus de filtrage appliqué aux configurations du quotient q encodé sur $L_0=4$ bits.
Pour éliminer les valeurs strictement supérieures à $q^{(0)}_{\max}=1011_2$, nous définissons $J^{(0)}_{\max}=\{3,1,0\}$ et $\mathcal{B}^{(0)}_{\max}=\{3\}$. Nous introduisons une variable auxiliaire $p^{\max}_{0,3}$ en appliquant la contrainte (6a), représentant le préfixe de $q^{(0)}_{\max}$ jusqu'à l'indice 3, ce qui donne les clauses suivantes : $(\bar{p}^{\max}_{0,3}\vee x_{0,3})$ ainsi que $(p^{\max}_{0,3}\vee\bar{x}_{0,3})$. Ensuite, en appliquant la contrainte (6c), nous imposons que la variable $x_{0,2}$ ne puisse pas prendre la valeur 0 lorsque $p^{\max}_{0,3}$ est vraie, conduisant à la clause $(\bar{p}^{\max}_{0,3}\vee x_{0,2})$. Nous procédons de manière analogue pour éliminer les valeurs strictement inférieures à $q^{(0)}_{\min}=0101_2$.

4 Évaluation expérimentale

4.1 Protocole expérimental

Pour évaluer les modèles introduits dans la section précédente, nous avons testé des instances issues du Large Weight Ternary Syndrome Decoding Challenge[1] (LW3SD), avec des valeurs de n allant de 10 à 100. Nous désignons nos modèles (2), (3), (4;5), et (4;6) respectivement par CNF1, CNF2, CNF3 et CNF4. De plus, nous avons implémenté ces modèles à l'aide du solveur CP-SAT

1. https://decodingchallenge.org/large-weight

Instance		CaDiCaL				CryptoMiniSat				CP-SAT			
n	t	CNF1	CNF2	CNF3	CNF4	CNF1	CNF2	CNF3	CNF4	CP1	CP2	CP3	CP4
10	9	89.26	14.05	78.33	26.47	86.21	15.06	80.4	26.3	**0.44**	0.57	0.55	**0.44**
20	19	49.52	192.04	1.02	1.13	38.01	200.92	10.88	3.02	0.59	0.7	0.55	**0.5**
30	29	17.85	147.73	63.55	**2.09**	6.98	31.64	152.1	66.58	3.20	9.24	2.31	1.89
40	39	52.15	103.14	115.35	20.14	23.91	126.5	26.26	67.19	16.86	33.23	14.71	**11.81**
50	49	90.41	71.71	**69.14**	87.7	173.34	250.39	79.73	145.62	161.79	328.22	82.74	85.5
60	59	807.09	**567.04**	577.33	852.62	1166.52	980.41	782.54	1107.86	1614.22	3806.92	807.07	1300.56
70	69	11772.57	14143.43	**6583.66**	17734.06	11165.61	10752.38	9054.9	13356.76	20704.41 (11)	10295.93 (6)	25763.99 (14)	23997.54 (15)
80	79	7888.75 (4)	10527.77 (7)	6855.34 (3)	8636.05 (7)	14877.03 (11)	**17253.96 (12)**	19644.53 (10)	9592.45 (5)	851.66 (1)	- (0)	- (0)	6984.33 (5)
90	89	**4580.82 (2)**	- (0)	- (0)	- (0)	464.74 (1)	- (0)	1692.33 (1)	2709.72 (1)	- (0)	- (0)	- (0)	2552.23 (1)
100	99	**533.04 (1)**	- (0)	- (0)	- (0)	802.53 (1)	- (0)	- (0)	- (0)	- (0)	- (0)	2309.31 (1)	- (0)
TOTAL		25881.48 (147)	25766.92 (147)	14343.72 (143)	27360.25 (147)	**28804.88 (153)**	29611.25 (152)	31523.67 (151)	27075.5 (146)	23353.18 (132)	14474.82 (126)	28981.23 (135)	34934.8 (141)

TABLE 2 – Résultats en termes de temps de résolution pour chaque taille n et poids t pour les différents modèles du problème LW3SD. Si les problèmes n'ont pas été résolus dans le temps imparti, le nombre d'instances résolues est indiqué entre parenthèses "()". Les meilleurs résultats sont mis en évidence en gras, d'abord sur la base du nombre d'instances résolues, puis, en cas d'égalité, sur la base du temps de résolution.

d'OR-Tools[2], notés CP1, CP2, CP3 et CP4, correspondant à leurs formulations CNF respectives. Les expérimentations ont été menées en utilisant les solveurs SAT CryptoMiniSat [30] et CaDiCaL [7], ainsi que le solveur CP-SAT. Tous les tests ont été effectués sur la plateforme de calcul MatriCS[3], sur la partition "bigmem" équipée de 12 serveurs biprocesseurs Intel Xeon E5-2680 v4 (2,40 GHz) disposant chacun de 512 Go de mémoire, avec une limite de temps fixée à 3600 secondes (1 heure) par instance.

Nous avons utilisé la bibliothèque PySAT[4] pour encoder nos modèles ainsi que les contraintes Pseudo-Booléennes. En particulier, nous utilisons l'encodage CardNetwork [1] pour les contraintes de cardinalité présentes dans tous les modèles CNF. Pour une contrainte de cardinalité de la forme $\sum_{i=1}^{h} l_i \leq b$, la complexité en termes de nombre de clauses (et de variables) générées est de $O(h \log^2 b)$ pour CardNetwork. Nous avons choisi l'encodage BinMerge [18] pour encoder les contraintes Pseudo-Booléennes, qui présente une complexité de $O(h \log^2 h \log W)$, où h désigne le nombre de littéraux dans la somme et W le poids maximal. Enfin, pour chaque valeur de n et chaque modèle, nous avons généré 20 instances en utilisant des graines aléatoires distinctes. Notre code source pour la génération des instances est disponible en ligne[5].

4.2 Résultats

Les résultats concernant le nombre d'instances résolues et les temps de résolution sont présentés dans le Tableau 2. Cette analyse met en évidence une distinction nette entre efficacité immédiate et passage à l'échelle. Sur les instances de petite taille, le solveur CP-SAT, et en particulier le modèle CP4, domine en termes de rapidité. Cependant, cette efficacité initiale ne se maintient pas lorsque la taille des instances augmente : les modèles CP voient leurs performances se dégrader fortement à partir de $n \geq 70$.

À l'inverse, les solveurs CaDiCaL et CryptoMiniSat se montrent plus robustes sur les instances de grande taille. CryptoMiniSat, associé au modèle CNF1, obtient la meilleure performance globale avec un total de 153 instances résolues. Bien que moins rapides sur les cas les plus simples, ces solveurs restent capables de traiter des instances de taille $n \geq 70$, là où les approches CP échouent quasi systématiquement à produire une solution dans le temps imparti.

L'analyse du compromis entre temps de résolution et taux de succès met en avant le modèles CNF1 associé à CryptoMiniSAT. Contrairement à CP-SAT, dont les performances se dégradent fortement sur les grandes instances, jusqu'à ne résoudre presque aucune instance pour $n \geq 80$, les approches SAT maintiennent une progression bien plus prévisible et stable. Le modèle CNF1 s'impose ici comme le plus robuste, étant le seul modèle capable de résoudre des instances jusqu'à $n = 100$.

La Figure 1 présente les temps de résolution triés par ordre croissant pour chaque configuration. Les courbes permettent de visualiser clairement le passage à l'échelle de chaque approche. On observe que les configurations CP-SAT décollent très tôt : leurs courbes commencent à diverger fortement à partir des instances $n = 120$, avec des temps qui atteignent rapidement le plafond de 3600 secondes. À l'inverse, les courbes SAT restent nettement plus basses sur l'ensemble de la distribution, ne dépassant le seuil de 1000 secondes que pour les instances au-delà de $n = 135$. Ce comportement illustre bien l'écart de passage à l'échelle : là où CP-SAT sature sur les instances difficiles, les solveurs SAT continuent à résoudre des instances que CP-SAT ne parvient pas à traiter.

Compte tenu de ces résultats, bien que CP-SAT s'avère efficace sur les instances de petite taille, il ne parvient pas à maintenir des performances compétitives à mesure que la difficulté du problème augmente, résolvant nettement moins d'instances que les configurations SAT. En raison de cet écart marqué en termes de passage à l'échelle et de taux de réussite global, la suite de cette étude se concentrera exclusivement sur l'analyse des solveurs SAT, car ils démontrent une capacité supérieure à traiter les instances les plus complexes du problème LW3SD.

2. https://developers.google.com/optimization/cp/cp_solver
3. https://www.matrics.u-picardie.fr/
4. https://pysathq.github.io/
5. https://github.com/carlberton/sat-LWZSD.git

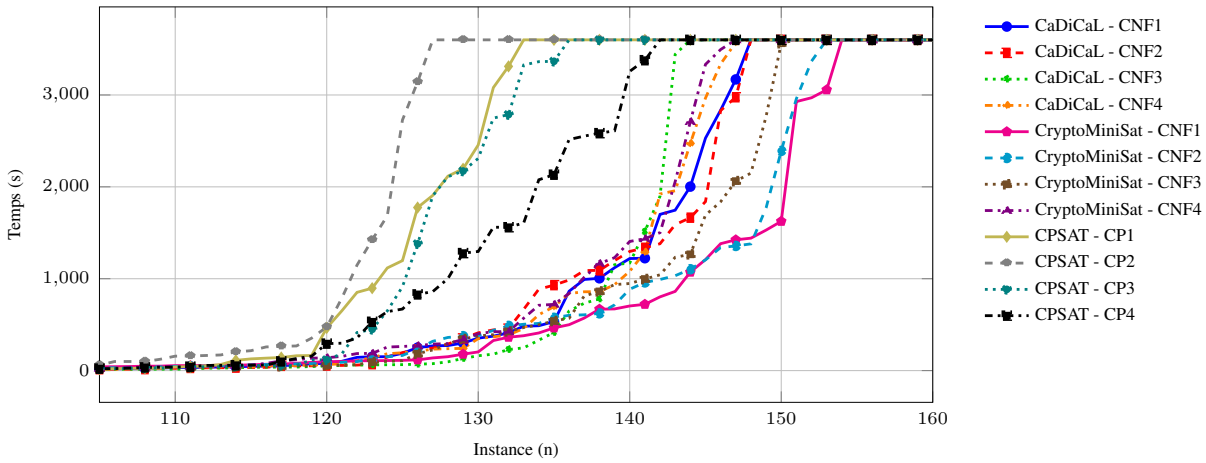


FIGURE 1 – Évolution des temps de résolution triés par ordre croissant pour les différents solveurs et modèles du problème LW3SD.

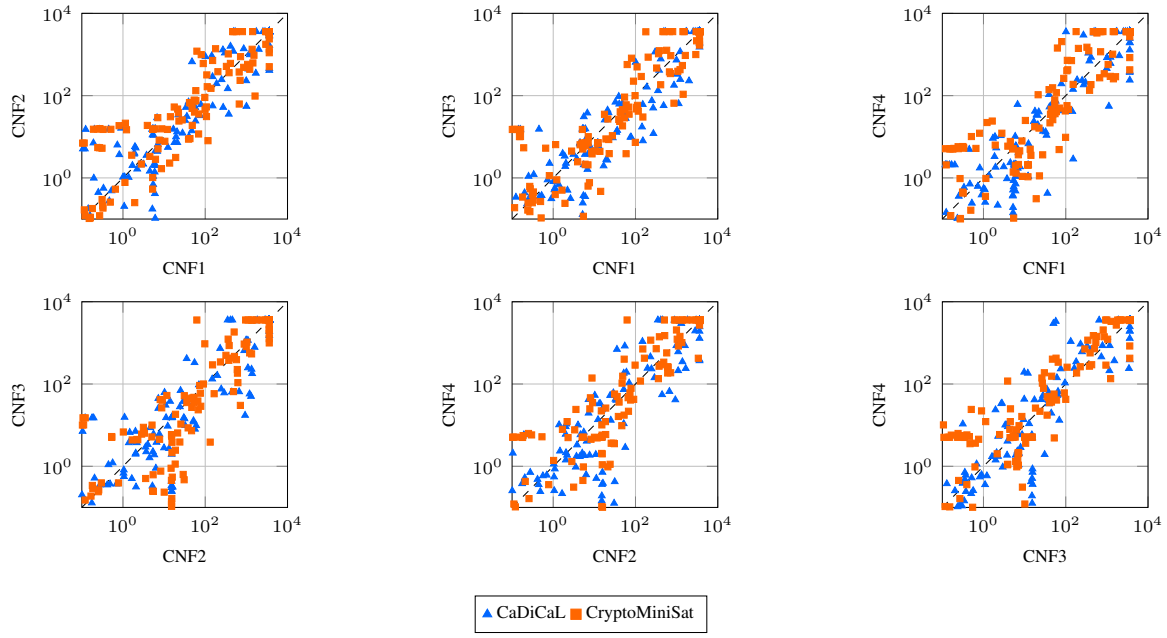


FIGURE 2 – Comparaison des temps de résolution en secondes par instance du problème LW3SD.

La Figure 2 présente une comparaison par instance sous forme de nuage de points comparatif de paires de modèles. L'encodage CNF1 s'impose comme la solution la plus robuste : il surpasse CNF4 dans 48% des cas avec CaDiCaL contre seulement 29% pour CNF4. Cette domination se confirme face à CNF2, où CNF1 l'emporte dans 41.5% des instances avec CryptoMiniSat. À l'inverse, l'encodage CNF3 s'avère être le moins efficace : il est systématiquement surpassé par CNF4, qui s'impose dans 43% des instances contre seulement 34% pour CNF3 avec CryptoMiniSat, validant ainsi l'intérêt de l'approche par filtrage compact par rapport au filtrage exhaustif. Enfin, la comparaison entre CNF2 et CNF4 montre des performances équilibrées, bien qu'un léger avantage se dessine pour CNF2 avec CryptoMiniSat (38.5% de victoires contre 34%). Enfin, il est important de noter qu'environ 20% à 26% des instances se distinguent par une performance similaires, principalement due aux limites de temps atteintes par les deux configurations ou à des résolutions instantanées.

5 Conclusion

Cet article présente une étude de différentes modélisations CNF pour le problème du décodage de syndrome à poids élevé (LWZSD) sur le corps fini $\mathbb{F}_Z$ (avec Z premier). Nous avons mené une évaluation expérimentale sur des instances ternaires ($Z = 3$) en comparant les performances des modèles introduits à l'aide de solveurs SAT de l'état de l'art, plus précisément CryptoMiniSat et CaDiCaL, ainsi que du solveur CP-SAT. Nos expériences démontrent que les sol-

veurs SAT sont beaucoup plus robustes et parviennent à résoudre les instances les plus complexes.
Une perspective intéressante de nos travaux consiste à adapter les mécanismes internes des solveurs SAT afin de mieux exploiter la structure intrinsèque du problème LWZSD, notamment par la conception d'heuristiques de branchement dédiées. Par exemple, on pourrait favoriser les variables d'erreur en fonction de la distribution des coefficients dans les équations de parité ou du poids d'erreur cible. Par ailleurs, il serait intéressant de concevoir des modèles plus compacts capturant mieux les relations complexes entre les différentes équations du syndrome. De tels modèles pourraient exploiter les similitudes structurelles apparaissant entre les équations et éviter ainsi d'encoder des motifs redondants. Cela permettrait potentiellement de réduire la taille de l'encodage et d'améliorer l'efficacité des solveurs. Enfin, pour traiter des instances encore plus complexes, il serait pertinent d'explorer des méthodes incomplètes dédiées ou de développer des approches hybrides combinant techniques complètes et incomplètes, afin de concilier le passage à l'échelle avec les garanties de solution.

Remerciements

Ce travail est soutenu par les projets ANR-24-CE23-6126 (BforSAT) et 22-PECY-0010 (CRYPTANALYSE), financés par l'Agence Nationale de la Recherche (ANR). Ce travail a également bénéficié d'un accès aux ressources de calcul de la "Plateforme MatriCS" de l'Université de Picardie Jules Verne, cofinancée par l'Union européenne via le Fonds Européen de Développement Régional (FEDER) et le Conseil Régional des Hauts-de-France.

Références

[1] Roberto Asín, Robert Nieuwenhuis, Albert Oliveras, and Enric Rodríguez-Carbonell. Cardinality networks and their applications. In Oliver Kullmann, editor, *SAT 2009*, volume 5584 of *LNCS*, pages 167–180. Springer, 2009.

[2] Anja Becker, Antoine Joux, Alexander May, and Alexander Meurer. Decoding random binary linear codes in 2 n/20 : How 1 + 1 = 0 improves information set decoding. In David Pointcheval and Thomas Johansson, editors, *Advances in Cryptology - EUROCRYPT 2012 - 31st Annual International Conference on the Theory and Applications of Cryptographic Techniques, Cambridge, UK, April 15-19, 2012. Proceedings*, volume 7237 of *Lecture Notes in Computer Science*, pages 520–536. Springer, 2012.

[3] Elwyn R. Berlekamp. *Algebraic coding theory*. McGraw-Hill series in systems science. McGraw-Hill, 1968.

[4] Elwyn R. Berlekamp, Robert J. McEliece, and Henk C. A. van Tilborg. On the inherent intractability of certain coding problems (corresp.). *IEEE Trans. Inf. Theory*, 24(3) :384–386, 1978.

[5] Carl Berton, Sami Cherif, and Claire Delaplace. Satisfiabilité pour le décodage par syndrome. In *Journées Francophones de Programmation par Contraintes (JFPC 2025)*, Dijon, France, June 2025.

[6] Carl Berton, Sami Cherif, and Claire Delaplace. Sat-based syndrome decoding and low-weight codewords. In *27th International Symposium on Formal Methods (FM 2026)*, Tokyo, Japan, 2026. To appear.

[7] Armin Biere, Tobias Faller, Katalin Fazekas, Mathias Fleury, Nils Froleyks, and Florian Pollitt. Cadical 2.0. In Arie Gurfinkel and Vijay Ganesh, editors, *Computer Aided Verification - 36th International Conference, CAV 2024, Montreal, QC, Canada, July 24-27, 2024, Proceedings, Part I*, volume 14681 of *Lecture Notes in Computer Science*, pages 133–152. Springer, 2024.

[8] Armin Biere, Marijn Heule, Hans van Maaren, and Toby Walsh, editors. *Handbook of Satisfiability - Second Edition*, volume 336 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2021.

[9] Rémi Bricout, André Chailloux, Thomas Debris-Alazard, and Matthieu Lequesne. Ternary syndrome decoding with large weight. In Kenneth G. Paterson and Douglas Stebila, editors, *Selected Areas in Cryptography - SAC 2019 - 26th International Conference, Waterloo, ON, Canada, August 12-16, 2019, Revised Selected Papers*, volume 11959 of *Lecture Notes in Computer Science*, pages 437–466. Springer, 2019.

[10] Lily Chen, Lily Chen, Stephen Jordan, Yi-Kai Liu, Dustin Moody, Rene Peralta, Ray A Perlner, and Daniel Smith-Tone. *Report on post-quantum cryptography*, volume 12. US Department of Commerce, National Institute of Standards and Technology . . ., 2016.

[11] Stephen A. Cook. The complexity of theorem-proving procedures. In Michael A. Harrison, Ranan B. Banerji, and Jeffrey D. Ullman, editors, *STOC 1971*, pages 151–158. ACM, 1971.

[12] Nicolas T. Courtois and Gregory V. Bard. Algebraic cryptanalysis of the data encryption standard. In Steven D. Galbraith, editor, *Cryptography and Coding, 11th IMA International Conference, Cirencester, UK, December 18-20, 2007, Proceedings*, volume 4887 of *Lecture Notes in Computer Science*, pages 152–169. Springer, 2007.

[13] Joan Daemen and Vincent Rijmen. *The Design of Rijndael : AES - The Advanced Encryption Standard*. Information Security and Cryptography. Springer, 2002.

[14] Thomas Debris-Alazard, Nicolas Sendrier, and Jean-Pierre Tillich. Wave : A new code-based signature scheme. *IACR Cryptol. ePrint Arch.*, page 996, 2018.

[15] Andre Esser and Emanuele Bellini. Syndrome decoding estimator. In Goichiro Hanaoka, Junji Shikata, and Yohei Watanabe, editors, *IACR 2022*, volume 13177 of *LNCS*, pages 112–141. Springer, 2022.

[16] Aarti Gupta, Malay K. Ganai, and Chao Wang. Sat-based verification methods and applications in hardware verification. In Marco Bernardo and Alessandro Cimatti, editors, *SFM 2006, Advanced Lectures*, volume 3965 of *LNCS*, pages 108–143. Springer, 2006.

[17] Frédéric Lafitte, Jorge Nakahara Jr., and Dirk Van Heule. Applications of SAT solvers in cryptanalysis : Finding weak keys and preimages. *J. Satisf. Boolean Model. Comput.*, 9(1) :1–25, 2014.

[18] Norbert Manthey, Tobias Philipp, and Peter Steinke. A more compact translation of pseudo-boolean constraints into CNF such that generalized arc consistency is maintained. In Carsten Lutz and Michael Thielscher, editors, *KI 2014*, volume 8736 of *LNCS*, pages 123–134. Springer, 2014.

[19] R. J. McEliece. A Public-Key Cryptosystem Based On Algebraic Coding Theory. *Deep Space Network Progress Report*, 44 :114–116, January 1978.

[20] Carlos Aguilar Melchor, Nicolas Aragon, Slim Bettaieb, Loıc Bidoux, Olivier Blazy, Jean-Christophe Deneuville, Philippe Gaborit, Edoardo Persichetti, Gilles Zémor, and IC Bourges. Hamming quasi-cyclic (hqc). *NIST PQC Round*, 2(4) :13, 2018.

[21] Ilya Mironov and Lintao Zhang. Applications of SAT solvers to cryptanalysis of hash functions. In Armin Biere and Carla P. Gomes, editors, *Theory and Applications of Satisfiability Testing - SAT 2006, 9th International Conference, Seattle, WA, USA, August 12-15, 2006, Proceedings*, volume 4121 of *Lecture Notes in Computer Science*, pages 102–115. Springer, 2006.

[22] Shintaro Narisada, Shusaku Uemura, Hiroki Okada, Hiroki Furue, Yusuke Aikawa, and Kazuhide Fukushima. Solving mceliece-1409 in one day - cryptanalysis with the improved BJMM algorithm. In Nicky Mouha and Nick Nikiforakis, editors, *ISC 2024*, volume 15258 of *LNCS*, pages 3–23. Springer, 2024.

[23] Eugene Prange. The use of information sets in decoding cyclic codes. *IRE Trans. Inf. Theory*, 8(5) :5–9, 1962.

[24] Jussi Rintanen. Planning and SAT. In Armin Biere, Marijn Heule, Hans van Maaren, and Toby Walsh, editors, *Handbook of Satisfiability - Second Edition*, volume 336 of *Frontiers in Artificial Intelligence and Applications*, pages 765–789. IOS Press, 2021.

[25] Olivier Roussel and Vasco Manquinho. Pseudo-boolean and cardinality constraints. In Armin Biere, Marijn Heule, Hans van Maaren, and Toby Walsh, editors, *Handbook of Satisfiability - Second Edition*, Frontiers in Artificial Intelligence and Applications, pages 1087–1129. IOS Press, 2021.

[26] Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3) :379–423, 1948.

[27] Peter W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Rev.*, 41(2) :303–332, 1999.

[28] João P. Marques Silva and Karem A. Sakallah. GRASP - a new search algorithm for satisfiability. In Rob A. Rutenbar and Ralph H. J. M. Otten, editors, *ICCAD 1996*, pages 220–227. IEEE Computer Society / ACM, 1996.

[29] Mate Soos, Karsten Nohl, and Claude Castelluccia. Extending SAT solvers to cryptographic problems. In Oliver Kullmann, editor, *SAT 2009*, volume 5584 of *LNCS*, pages 244–257. Springer, 2009.

[30] Mate Soos, Karsten Nohl, and Claude Castelluccia. Extending SAT solvers to cryptographic problems. In Oliver Kullmann, editor, *SAT 2009*, volume 5584 of *LNCS*, pages 244–257. Springer, 2009.

[31] Jacques Stern. A method for finding codewords of small weight. In Gérard D. Cohen and Jacques Wolfmann, editors, *Coding Theory and Applications, 3rd International Colloquium, Toulon, France, November 2-4, 1988, Proceedings*, volume 388 of *Lecture Notes in Computer Science*, pages 106–113. Springer, 1988.

[32] Monika Trimoska, Gilles Dequen, and Sorina Ionica. Logical cryptanalysis with wdsat. In Chu-Min Li and Felip Manyà, editors, *Theory and Applications of Satisfiability Testing - SAT 2021 - 24th International Conference, Barcelona, Spain, July 5-9, 2021, Proceedings*, volume 12831 of *Lecture Notes in Computer Science*, pages 545–561. Springer, 2021.

Une approche SAT pour l'analyse de l'autostabilisation d'un unisson asynchrone

Asma Khoualdia, Sami Cherif, Stéphane Devismes, Léo Robert

Laboratoire MIS UR 4290, Université de Picardie Jules Verne, Amiens, France

{prenom.nom}@u-picardie.fr

Résumé

L'unisson asynchrone est un problème de synchronisation d'horloges de référence en systèmes distribués, et notamment en autostabilisation. Il permet notamment de simuler des systèmes synchrones dans des environnements asynchrones, faciliter la détection de terminaison, partager localement des ressources, ou encore pour réaliser des calculs d'infimum. Cet article explore l'autostabilisation d'un algorithme d'unisson asynchrone en utilisant une approche basée sur la satisfiabilité propositionnelle. Nous proposons une modélisation logique de l'exécution asynchrone de l'algorithme capturant les règles de mise à jour des horloges ainsi que la détection de la convergence vers une configuration légitime ou, au contraire, de la divergence par l'existence d'interblocages ou de cycles entre configurations illégitimes.

Mots-clés

Autostabilisation, unisson, asynchronisme, satisfiabilité.

Abstract

Asynchronous unison is a benchmark clock synchronization problem in distributed systems and particularly self-stabilization. It allows to simulate a synchronous system in an asynchronous environment, help the termination detection, achieve local resource allocation, or compute infimum. This article explores the self-stabilization of an asynchronous unison algorithm using a propositional satisfiability-based approach. We propose a logical model of the asynchronous execution of this algorithm capturing the clock update rules, as well as the detection of convergence to a legitimate configuration or, conversely, divergence through the existence of daedlocks or cycles of illegitimate configurations.

Keywords

Self-stabilization, Unison, Asynchronism, Satisfiability.

1 Introduction

Un système distribué est un ensemble d'entités de calcul, ici appelées *processus*, qui sont *autonomes* et *interconnectées* entre elles. Le but de ces processus est de coopérer, via des échanges d'informations, afin de résoudre une tâche globale au système. Dans ce contexte, un algorithme est dit *autostabilisant* si, quel que soit l'état initial du système sur lequel il est déployé, il garantit le retour en temps fini, et sans intervention extérieure, à un état dit *légitime* à partir duquel la spécification du système est satisfaite [16]. Bien que la notion de panne ne soit pas explicite dans la définition des systèmes autostabilisants, la motivation première de cette approche est bien la *tolérance aux pannes*. En effet, après un nombre fini de *pannes transitoires* [1], l'état d'un système distribué peut être quelconque et ainsi ne plus vérifier sa spécification. Un algorithme autostabilisant garantit alors que le système récupère de telles fautes en temps fini.

La conception d'algorithmes autostabilisants peut sembler complexe au premier abord en raison de l'accès limité des processus à l'état global du système. Pourtant, nombre de ces algorithmes se révèlent étonnamment élégants et parfois même plus simples que leurs pendants non stabilisants. Cependant, démontrer formellement l'autostabilisation d'un algorithme s'avère souvent ardu et sujet aux erreurs. Cela est principalement dû à la nature combinatoire du problème : la convergence vers un état légitime doit être démontrée à partir de *tout* état du système.

Dans cette optique, nous nous intéressons ici à un problème simple : l'*unisson asynchrone*. On suppose un système asynchrone où les processus sont anonymes et arbitrairement connectés entre-eux. Chaque processus dispose d'une horloge locale dont les valeurs entières varient entre 0 à $K-1$ où K est appelée la *période*. À partir d'un état où l'horloge de chaque processus a une valeur quelconque de $\{0,\ldots,K-1\}$, le but est de faire converger le système vers un état dans lequel la différence entre les horloges de deux voisins quelconques est d'au plus un incrément (modulo K) à chaque instant. L'unisson est un outil d'algorithmique distribuée fondamental qui possède de nombreuses applications. Entre autres, il peut être utilisé pour simuler des systèmes synchrones dans des environnements asynchrones [15], pour faciliter la détection de terminaison [8], pour partager localement des ressources [10], ou encore pour réaliser des calculs d'infimum [9].

Couvreur *et al.* ont proposé une solution simple à ce problème [13]. L'autostabilisation de leur algorithme est démontrée pour tout $K > n^2$ où n est la taille du réseau. Cependant, l'optimalité de cette borne reste une question

1. C'est-à-dire, des perturbations temporaires de certains composants du système (liens ou processus).

ouverte. Nous proposons ici d'étudier comment la satisfiabilité propositionnelle peut aider à traiter cette question, et plus généralement comment elle peut aider à vérifier l'autostabilisation d'algorithmes distribués asynchrones. En effet, la *satisfiabilité* propositionnelle (SAT) est un problème fondamental en logique et en informatique théorique, consistant à déterminer si une formule booléenne peut être satisfaite par une assignation de valeurs. Ce problème, qui a été le premier démontré NP-complet [12], suscite un intérêt croissant car les solveurs SAT modernes parviennent à traiter efficacement de grandes instances issues d'applications concrètes, malgré la difficulté théorique établie du problème. En plus de son rôle clé en vérification logicielle et matérielle, en intelligence artificielle et en cryptographie, SAT se situe à l'intersection de plusieurs disciplines telles que la logique, la complexité et la programmation par contraintes [4].

Dans [18], nous avons initié l'analyse rigoureuse d'algorithmes autostabilisants déployés dans des réseaux de topologies variées via la satisfiabilité propositionnelle en prenant l'unisson synchrone d'Arora *et al.* [2] comme cas d'étude. Cependant, l'étude proposée se restreint aux systèmes distribués synchrones. Dans cet article, nous proposons une approche formelle basée sur SAT pour analyser rigoureusement le comportement de l'unisson asynchrone [13]. Cet algorithme suit des règles de mise à jour d'horloges distinctes, et son autostabilisation se caractérise par des propriétés différentes. Nous traduisons ces règles en contraintes logiques, permettant l'utilisation de solveurs SAT pour vérifier si la synchronisation est atteinte dans un cadre asynchrone. Cette méthodologie offre un cadre systématique pour étudier l'autostabilisation de l'algorithme dans différentes topologies et configurations initiales et avec différentes périodes.

Le reste de cet article est organisé comme suit. Dans la section 2, nous présentons l'algorithme d'unisson asynchrone et nous introduisons la satisfiabilité propositionnelle. La section 3 est consacrée à la modélisation formelle de l'algorithme de l'unisson asynchrone, où nous détaillons les variables et les contraintes logiques, avec une attention portée à la mise à jour des horloges et à la modélisation des propriétés d'autostabilisation. La section 4 présente les résultats obtenus. Enfin, nous concluons dans la section 5.

2 Préliminaires

2.1 Unisson asynchrone

Nous considérons un système distribué de $n \geq 3$ processus interconnectés. Le réseau d'interconnexion est modélisé par un graphe non-orienté connexe $G = (V, E)$, où V est un ensemble de n sommets représentant les processus et E un ensemble d'arêtes représentant les liens de communication bidirectionnels entre les processus. L'ensemble des voisins d'un processus p est noté $N(p)$. Nous considérons l'unisson asynchrone proposé par Couvreur *et al.* [13]. Cet algorithme est décrit dans le *modèle à états* dans [1]. Il s'agit d'un modèle à mémoires localement partagées : chaque nœud peut directement lire son état et ceux de ses voisins, mais peut uniquement modifier son propre état.

Algorithme 1 Unisson asynchrone : algorithme local pour chaque processus p

Entrées :

$N(p)$	:	l'ensemble des voisins de p
K	:	entier strictement positif

Variable :

$p.c \in \{0, \ldots, K-1\}$

Prédicat :

$behind(a, b) = ((b.c - a.c) \mod K) \leq n$

Règles :

$I(p) :: \forall q \in N(p),\ behind(p, q) \rightarrow p.c \leftarrow (p.c + 1) \bmod K$

$R(p) :: p.c \neq 0 \wedge (\exists q \in N(p),\ \neg behind(p, q) \wedge \neg behind(q, p)) \rightarrow p.c \leftarrow 0$

Étant donné un paramètre entier K commun à tous les processus et appelé *période*, chaque processus p dispose d'une seule variable $p.c \in \{0, \ldots, K-1\}$ qui représente son *horloge*. La *configuration* du système est ainsi modélisée par le vecteur associant à chaque processus la valeur de son horloge. L'ensemble des configurations possibles du système sera noté $\Gamma_{n,K}$. L'algorithme local de chaque processus p est décrit par deux règles de la forme $étiquette :: garde \rightarrow action$. L'étiquette est uniquement utilisée pour identifier la règle dans le raisonnement. La garde d'une règle de p est un prédicat sur l'état de p et ceux de ses voisins. L'action correspond à une mise-à-jour de la valeur d'horloge du processus p. Dans une configuration donnée, la règle d'un processus est dite activable si sa garde est vraie. Par extension, un processus est dit activable lorsqu'une de ses règles est activable. Une configuration où aucune règle n'est activable est dite *terminale*.

Le système est asynchrone : tant que le système n'est pas dans une configuration terminale, des *itérations asynchrones* sont exécutées comme suit. Un adversaire, appelé *démon distribué inéquitable*[2], sélectionne un sous-ensemble non vide de processus activables dans la configuration courante γ_i. Les processus sélectionnés sont dits *activés* et exécutent simultanément l'action de l'une de leurs règles activables, générant ainsi une nouvelle configuration γ_{i+1}, et ainsi de suite. Une exécution de l'algorithme est donc une suite maximale (finie ou infinie) de configurations $e = \gamma_0, \gamma_1, \ldots$ telle que pour tout $i > 0$, γ_i est obtenu à partir de γ_{i-1} en une itération asynchrone de l'algorithme.

Dans l'algorithme 1, le prédicat $behind(p, q)$ est vrai lorsque le processus p est en retard d'au plus n incréments par rapport à q. On dira dans ce cas que p a un *retard admissible* par rapport à q. Ainsi, un processus p est activable pour l'incrémentation ($I(p)$) dans une configuration donnée s'il a un retard admissible avec chacun de ses voisins. Dans ce cas, si le processus est choisi par le démon, il incrémente sa valeur modulo K pour rattraper son retard. D'autre part, un processus p est activable pour la réinitialisation ($R(p)$)

2. Le démon matérialise l'asynchronisme du système.

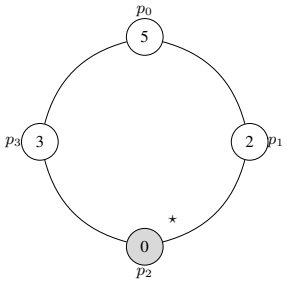


FIGURE 1 – Configuration initiale convergente γ_0 dans un anneau orienté de 4 processus avec une période $K = 9$. Un processus est grisé lorsqu'il est activable et affiche une étoile lorsqu'il est activé.

	p_0	p_1	p_2	p_3
γ_0	5	2	0*	3
γ_1	5	2	1*	3
γ_2	5	2*	2	3
γ_3	5	3	2*	3
γ_4	5	3*	3	3
γ_5	5	4	3*	3
γ_6	5	4*	4	3
γ_7	5	5	4	3*
γ_8	5	5	4	4

TABLE 1 – Valeurs des processus pour chaque configuration. Une case est grisée lorsque son processus est activable et affiche une étoile lorsqu'il est activé.

dans une configuration donnée si sa valeur d'horloge est non nulle et s'il a un voisin q avec qui son horloge n'est pas comparable : p n'a pas de retard admissible avec q et q n'a pas de retard admissible avec p. Dans ce cas, p est activable pour réinitialiser sa valeur d'horloge à 0 afin de sortir de cette situation ambiguë avec q. On note que les deux propriétés d'autorisation de l'algorithme 1 sont localement mutuellement exclusive : lorsqu'une des deux règles est activable pour un processus, l'autre ne l'est pas pour ce même processus. Ainsi, au plus une règle peut être applicable pour un processus à un instant donné.

Dans la suite, nous nous intéressons aux propriétés d'autostabilisation de l'algorithme. En particulier, on parle de convergence, définie formellement ci-dessous, quand on garantit un retour à une configuration légitime de l'algorithme où, pour chaque paire de voisins, la différence entre leurs valeurs d'horloge est d'au plus un incrément. Précisément, dans une configuration légitime, on a $p.c \in \{q.c - 1 \bmod K, q.c, q.c + 1 \bmod K\}$, pour toute paire de voisins p et q. À noter ici que, les horloges étant modulaires, l'écart entre $K - 1$ et 0 n'est que d'un seul incrément. De plus, l'ensemble des configurations légitimes est clos : toute configuration accessible depuis une configuration légitime est également légitime.

L'algorithme est démontré convergent (ou autostabilisant) pour toute période $K > n^2$ [14]. Cependant, lorsque la valeur de K est inférieure ou égale à n^2, l'exécution peut ne jamais stabiliser. On parle alors de *divergence*, définie formellement ci-dessous, quand un *cycle de configurations*

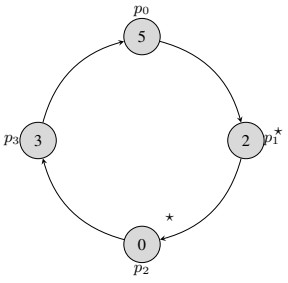


FIGURE 2 – Configuration initiale divergente γ_0 par l'existence d'un cycle (de configurations illégitimes) dans un anneau de 4 processus avec $K = 6$. Un processus est grisé lorsqu'il est activable et affiche une étoile lorsqu'il est activé.

	p_0	p_1	p_2	p_3
γ_0	5	2*	0*	3
γ_1	5*	3*	1*	3*
γ_2	0*	4*	2	4*
γ_3	1	5	2*	5
γ_4	1*	5	3	5
γ_5	2	5*	3	5*
γ_6	2*	0	3*	0*
γ_7	3*	0*	4*	1*
γ_8	4*	1*	5*	2*
γ_9	5	2	0	3

TABLE 2 – Valeurs des processus pour chaque configuration avec quatre processus. Une case est grisée lorsque le processus est activable et une étoile indique le processus effectivement activé.

illégitimes est exhibé dans un préfixe d'exécutions ou lorsqu'une exécution atteint un *interblocage*, c'est-à-dire, une configuration terminale (illégitime). Un cycle peut être répété à l'infini et ainsi donner une exécution ne contenant aucune configuration légitime. Une configuration terminale viole la spécification de l'unisson, qui exige que chaque processus incrémente son horloge infiniment souvent. On illustre ces différents cas de figures dans les exemples 1, 2 et 3.

Définition 1 (Convergence)**.** *L'algorithme de l'unisson asynchrone exécuté sur un graphe non orienté connexe de n processus est dit* convergent *si, à partir de toute configuration initiale, toutes les exécutions possibles atteignent en un nombre fini d'itérations une configuration* légitime *où, pour chaque paire de voisins, la différence entre leurs valeurs d'horloges est au plus un incrément.*

Exemple 1. *Considérons un réseau de processus G=($\{p_0, p_1, p_2, p_3\}, E$) dont la topologie est en anneau. Posons $K = 9$ et supposons la configuration initiale γ_0 donnée dans la Figure 1. À partir de γ_0, on obtient le préfixe d'exécution $\gamma_0, \gamma_1, \ldots, \gamma_8$ donné dans la Table 1. Après la dernière itération, l'algorithme atteint une configuration λ_8 légitime puisqu'en particulier, toutes les valeurs d'horloges ont un écart d'au plus 1.*

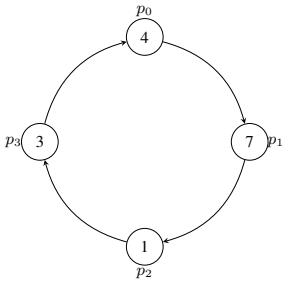


FIGURE 3 – Configuration terminale divergente γ_0 par l'existence d'un interblocage dans un anneau de 4 processus avec une période $K = 9$.

Définition 2 (Divergence)**.** *L'algorithme de l'unisson asynchrone exécuté sur un graphe non orienté connexe de n processus est dit* divergent *s'il existe un interblocage (c'est-à-dire, une configuration terminale) ou une paire de configurations identiques (appelée cycle) dans une exécution asynchrone où, pour (au moins) une paire de voisins, la différence entre leurs valeurs d'horloge dépasse un incrément.*

Exemple 2. *Considérons un réseau de processus $G = (\{p_0, p_1, p_2, p_3\}, E)$ dont la topologie est en anneau. Posons $K = 6$ et supposons la configuration initiale γ_0 donnée dans la Figure 2. À partir de γ_0, on obtient le préfixe d'exécution $\gamma_0, \gamma_1, \ldots, \gamma_9$ détaillé dans la Table 2. Un cycle de configurations illégitimes apparaît à partir de la configuration γ_9.*

Exemple 3. *Considérons un réseau de processus $G = (\{p_0, p_1, p_2, p_3\}, E)$ dont la topologie est en anneau et posons $K = 9$. La configuration terminale γ_0 donnée dans la Figure 3 est divergente par interblockage puisqu'aucun processus ne peut être activé.*

2.2 Satisfiabilité propositionnelle

Soit X un ensemble de variables propositionnelles. Un littéral est une variable $x \in X$ ou sa négation $\overline{x}$. Une clause est une disjonction de littéraux. Une formule en Forme Normale Conjonctive (FNC) est une conjonction de clauses. Une affectation $I : X \to \{vrai, faux\}$ associe à chaque variable une valeur booléenne et peut être représentée comme un ensemble de littéraux. Un littéral l est satisfait par une affectation I si $l \in I$, sinon il est falsifié par I. Une clause est satisfaite par une affectation I si au moins un de ses littéraux est satisfait par I, sinon elle est falsifiée par I. Une formule FNC est satisfiable s'il existe une affectation qui satisfait toutes ses clauses, sinon elle est insatisfiable. Le problème de satisfiabilité (SAT) consiste à déterminer si une formule FNC donnée est satisfiable.
Bien que SAT soit NP-complet [12], les solveurs modernes basés sur l'algorithme CDCL (Conflict Driven Clause Learning) [19] sont efficaces et parviennent à résoudre des instances impliquant un grand nombre de variables et de clauses. En effet, au delà de l'apprentissage de clauses et du retour arrière non chronologique, ces solveurs intègrent des mécanismes puissants comme, par exemple, les structures paresseuses, des heuristiques de branchement dédiées ou encore les redémarrages [7]. Par ailleurs, le problème SAT est largement utilisé pour modéliser et résoudre de nombreux problèmes combinatoires complexes issus de divers domaines, notamment en vérification formelle [5, 11] mais aussi en cryptographie, en bio-informatique et en planification [7].
Dans la suite, on s'intéresse à la modélisation du comportement de l'algorithme d'unisson asynchrone, appliqué à un graphe connecté. L'objectif est de modéliser l'exécution de l'algorithme et de caractériser ses conditions de convergence ou de divergence, à travers des formules logiques pouvant être résolues par un solveur SAT.

3 Modélisation formelle d'un algorithme d'unisson asynchrone

On rappelle que le réseau est modélisé par un graphe non orienté connexe de $n \geq 3$ processus, noté $G = (V, E)$. L'ensemble des processus est défini, sans perte de généralité, par leurs indices $V = \{0, \ldots, n-1\}$. L'ensemble des processus voisins à $p \in V$ dans G est noté $N(p)$. Les valeurs d'horloge de chaque processus sont contenues dans l'intervalle $M = \{0, \ldots, K-1\}$, où K est la période. Nous nous intéressons aux t_f premières configurations d'une exécution quelconque de l'algorithme de l'unisson asynchrone. Ces configurations sont indexées de 0 à $t_f - 1$, et l'ensemble des indices est noté $T = \{0, \ldots, t_f - 1\}$. Dans la suite, nous nous intéressons aux contraintes qui encadrent l'exécution de l'algorithme. Ces contraintes doivent non seulement représenter fidèlement l'évolution du système dans un contexte asynchrone, mais aussi permettre de détecter sa convergence ou sa divergence.

3.1 Modélisation de l'exécution

Dans un premier temps, nous nous intéressons à la formalisation des contraintes modélisant une exécution valide de l'unisson asynchrone. Plus précisément, nous décrivons les variables booléennes permettant de raisonner sur l'évolution du système et nous décrivons les contraintes permettant de représenter les règles de mise à jour, assurant ainsi l'évolution correcte des horloges au fil des itérations.

3.1.1 Sur l'unicité des horloges

Pour représenter l'évolution du réseau au fil des itérations de l'algorithme, nous introduisons les **variables d'horloge** $clk_{p,t,v}$, où $p \in V$, $t \in T$ et $v \in M$, qui sont affectées à $vrai$ si le processus p possède la valeur v dans la configuration t. Ces variables permettent de représenter l'état du système à chaque étape de l'exécution. Ainsi, pour garantir une exécution valide de l'algorithme, il est nécessaire de maintenir l'unicité de l'horloge de chaque processus à chaque itération. Précisément, il faut veiller à ce que chaque processus possède une et une seule valeur d'horloge dans chaque configuration du système. Cette contrainte peut être modélisée comme suit :

$$\sum_{v \in M} clk_{p,t,v} = 1, \qquad \forall p \in V, \forall t \in T \tag{1}$$

On peut naturellement écrire cette contrainte sous forme clausale via un encodage binomial classique comme décrit ci-dessous, nécessitant $O(n*t_f*K^2)$ clauses mais des formulations plus compactes peuvent néanmoins être obtenues grâce à l'introduction de variables auxiliaires [20].

$$\bigwedge_{\substack{(v,v')\in M^2 \\ v<v'}} (\overline{clk_{i,t,v}} \vee \overline{clk_{p,t,v'}}), \quad \forall p \in V, \forall t \in T \tag{1a}$$

$$\bigvee_{v\in M} clk_{p,t,v}, \quad \forall p \in V, \forall t \in T \tag{1b}$$

3.1.2 Sur l'autorisation des processus

On s'intéresse ici à la sémantique d'autorisation de mise à jour des horloges des processus dans l'unisson asynchrone. Pour cela, nous considérons tout d'abord la modélisation des retards légitimes entre paires de processus $(p, q) \in V^2$ avec $p \neq q$, définie par le prédicat $behind(p, q) = ((q.c - p.c) \bmod K) \leq n$. Pour cela, nous introduisons les **variables de retard** $bhd_{p,q,t}$ pour $p, q \in V^2$ tels que $p \neq q$ et $t \in T$, indiquant que la valeur d'horloge du processus p est en retard d'au plus n incréments par rapport à celle du processus q dans la configuration t. De plus, afin de faciliter l'encodage FNC, nous introduisons des **variables de double horloge** $dclk_{p,q,t,v,u}$ représentant la conjonction des deux variables d'horloge correspondantes, c'est-à-dire le fait que p possède la valeur v et q la valeur u à l'instant t. Les contraintes suivantes expriment alors que $bhd_{p,q,t}$ est vraie si et seulement si l'une des paires (v, u) satisfait la condition de retard admissible, générant un nombre de clauses en $O(n^2 * t_f * K^2)$:

$$bhd_{p,q,t} \leftrightarrow \bigvee_{\substack{(v,u)\in M^2 \\ (u-v) \bmod K \leq n}} dclk_{p,q,t,v,u} \quad \forall p \in V, \forall q \in N(p), \forall t \in T \tag{3}$$

$$dclk_{p,q,t,v,u} \leftrightarrow clk_{p,t,v} \wedge clk_{q,t,u} \quad \forall p \in V, \forall q \in V \setminus \{p\}, \forall v \in M^2, \forall t \in T \tag{4}$$

On rappelle qu'un processus $p \in V$ est activable quand la garde d'une de ses règles est vérifiée et il peut donc exécuter son action s'il est choisi par le démon. L'algorithme de l'unisson asynchrone présente deux règles d'autorisation différentes. Tout d'abord, un processus p est autorisé à incrémenter son horloge si elle a au plus n incréments de retard sur les valeurs d'horloge de chacun de ses voisins. Pour représenter cette sémantique, nous introduisons les **variables d'autorisation à l'incrémentation** $inc_{p,t}$ où $p \in V$ et $t \in T$ qui seront affectées à $vrai$ si le processus p a un retard admissible par rapport à tous les processus dans son voisinage $N(p)$ à la configuration t. Cette condition peut alors être exprimée par la contrainte suivante générant $O(n^2 * t_f)$ clauses :

$$inc_{p,t} \leftrightarrow \bigwedge_{q\in N(p)} bhd_{p,q,t} \quad \forall p \in V, \forall t \in T \tag{5}$$

L'algorithme prévoit également une seconde règle permettant à un processus de réinitialiser sa valeur d'horloge lorsque sa situation par rapport à ses voisins devient ambiguë. Plus précisément, un processus p est autorisé à se réinitialiser si sa valeur d'horloge est non nulle et s'il existe au moins un voisin dont la valeur d'horloge n'est pas dans une relation de retard admissible avec celle de p. Pour représenter cette sémantique, nous introduisons les **variables d'autorisation à la réinitialisation** $res_{p,t}$ où $p \in V$ et $t \in T$, qui seront affectées à $vrai$ si le processus p est autorisé à réinitialiser son horloge à la configuration t. Cette condition repose sur l'existence d'un voisin q avec qui l'horloge de p n'est pas comparable. Pour cela, nous introduisons **les variables de double retard** $dbhd_{p,q,t}$ indiquant qu'au moins l'un des deux processus p et q est en retard admissible par rapport à l'autre. Ces conditions peuvent alors être exprimées par les contraintes suivantes générant $O(n^2 * t_f)$ clauses :

$$res_{p,t} \leftrightarrow \overline{clk_{p,t,0}} \wedge \Big(\bigvee_{q\in N(p)} \overline{dbhd_{p,q,t}} \Big) \quad \forall p \in V, \forall t \in T \tag{6}$$

$$dbhd_{p,q,t} \leftrightarrow bhd_{p,q,t} \vee bhd_{q,p,t} \quad \forall p \in V, \forall t \in T \tag{7}$$

3.1.3 Sur l'activation des processus

Dans cette section, nous nous intéressons à la prise en compte de l'asynchronisme, qui induit différentes exécutions possibles à partir de chaque configuration. Précisément, l'asynchronisme du système est généré par un démon distribué inéquitable, qui active à chaque étape un sous-ensemble arbitraire non vide de processus activables. Lorsqu'il est activé, un processus modifie sa valeur d'horloge conformément aux règles définies dans l'algorithme 1 selon son statut d'autorisation (incrémentation ou ré-initialisation). À l'instant t, notons A_t l'ensemble des processus activables. Sous un démon distribué inéquitable, n'importe quel sous-ensemble non vide de A_t peut être activé simultanément. Chaque choix de sous-ensemble correspond donc à une évolution possible du système à partir de la configuration t. Le nombre total d'exécutions possibles à cette étape est alors égal au nombre de sous-ensembles non vides de A_t, soit $2^{|A_t|} - 1$. La prise en compte conjointe de l'asynchronisme et de la diversité des configurations initiales induit donc une combinatoire particulièrement importante.

Pour modéliser la sémantique de l'asynchronisme, nous introduisons les **variables d'activation** $act_{p,t}$ où $p \in V$ et $t \in T$. Ces variables sont affectées à $vrai$ si le processus p est activé à la configuration t, et leur choix est laissé libre au solveur pour simuler les décisions d'activation d'un démon distribué inéquitable. Afin de garantir la cohérence du modèle, deux contraintes doivent être imposées : un processus activé doit être activable (pour l'incrémentation ou de la ré-initialisation de son horloge), et l'ensemble des processus activés à chaque étape doit être non vide s'il n'y a pas d'interblocage. En effet, il convient de souligner qu'un

blocage peut survenir au niveau des mécanismes d'autorisation. Dans certaines situations, le système peut atteindre un état dans lequel aucun processus n'est en mesure de progresser. Dans ce cas, on parle de phénomène d'interblocage (*deadlock*), le système entre alors dans un état de divergence, caractérisé par l'absence totale d'évolution du système (toute les valeurs d'horloge restent figées).
Dans le modèle décrit dans la section 2, lorsqu'un interblocage apparaît dans la configuration t d'une exécution, l'exécution termine : les configurations t' avec $t' > t$ ne sont pas définies. Or, dans notre modélisation, la plupart des variables sont définies pour toutes les valeurs $t \in [0..t_f - 1]$. Pour palier à ce problème, nous complémentons toutes exécutions atteignant un interblocage en t avec un suffixe constitué de $t_f - t - 1$ configurations identiques à la configuration t. Cette astuce nous permettra d'ailleurs de gérer l'existence d'un interblocage de manière analogue à l'existence d'un cycle de configurations illégitimes.
Pour modéliser une situation d'interblocage, nous introduisons les **variables d'interblocage** $dlock_t$ où $t \in T$. Ces variables sont affectées à $vrai$ si aucun processus n'est activable à la configuration t. Les trois contraintes sont exprimées ci-dessous, introduisant ainsi un nombre de clauses bornée en $O(n * t_f)$:

$$act_{p,t} \rightarrow (inc_{p,t} \vee res_{p,t}), \quad \forall p \in V,\ \forall t \in T \tag{8}$$

$$\overline{dlock_t} \rightarrow \bigvee_{p \in V} act_{p,t}, \quad \forall t \in T \tag{9}$$

$$dlock_t \leftrightarrow \bigwedge_{p \in V} (\overline{inc_{p,t}} \wedge \overline{res_{p,t}}), \quad \forall t \in T \tag{10}$$

3.1.4 Sur la mise à jour des horloges

On s'intéresse maintenant à la mise à jour des horloges des processus à chaque itération de l'algorithme. On rappelle que, dans un système asynchrone contrôlé par un démon distribué inéquitable, tous les processus activables ne sont pas nécessairement activés simultanément, et seuls certains d'entre eux sont autorisés à modifier leur état en fonction de la vérification des propriétés d'autorisation. Il est donc nécessaire de distinguer les situations possibles afin de modéliser fidèlement l'évolution des configurations du système. Plus précisément, à chaque étape t, la mise à jour des horloges dépend du statut de chaque processus, en fonction de son activation et de son autorisation. Un processus peut donc se trouver dans l'une des trois situation suivantes : (i) Il n'est pas activable et conserve donc son état courant ; (ii) Il est activable mais non activé et, dans ce cas, maintient également son état inchangé ; (iii) Il est activable et activé, auquel cas il met à jour sa valeur conformément aux deux règles de l'algorithme. Ces trois cas couvrent l'ensemble des comportements possibles et suivent donc des règles spécifiques qui sont détaillées ci-dessous.

(i) Processus non activables. Un processus est *non activable* à l'étape t lorsqu'il ne vérifie aucune propriété d'autorisation : il n'est autorisé ni à incrémenter sa valeur d'horloge ni à la ré-initialiser. Dans ce cas, son état doit rester inchangé entre les configurations t et $t + 1$ (en particulier, en cas d'interblocage). Autrement dit, si le processus $p \in V$ a la valeur v dans la configuration t et qu'il n'est activable pour aucune des deux propriétés d'autorisation ($\overline{inc_{p,t}} \wedge \overline{res_{p,t}}$), alors il maintient la valeur v dans la configuration $t + 1$. Cette contrainte est exprimée ci-dessous, introduisant un nombre de clauses en $O(n * t_f * K)$:

$$\begin{aligned} clk_{p,t,v} \wedge \overline{inc_{p,t}} \wedge \overline{res_{p,t}} \rightarrow clk_{p,t+1,v} \\ \forall p \in V,\ \forall t \in T \setminus \{t_f - 1\}\ \forall v \in V \end{aligned} \tag{11}$$

(ii) Processus activables et non activés. Un processus est *activable mais non activé* à l'étape t lorsqu'il est autorisé à changer de valeur mais qu'il n'est pas sélectionné par le démon pour exécuter une transition à cette étape. Dans ce cas, bien que le processus pourrait modifier sa valeur, il conserve son état actuel, c'est-à-dire que son état reste inchangé entre les étapes t et $t+1$. Autrement dit, si le processus $p \in V$ possède la valeur $v \in M$ à l'étape $t \in T \setminus t_f - 1$ et n'est pas activé ($act_{p,t} = faux$), alors il conservera cette valeur à l'étape suivante. Cette contrainte est représentée ci-dessous et induit un nombre de clauses borné en $O(n * t_f * K)$:

$$\begin{aligned} clk_{p,t,v} \wedge \overline{act_{p,t}} \rightarrow clk_{p,t+1,v}, \\ \forall p \in V,\ \forall t \in T \setminus \{t_f - 1\}\ , \forall v \in M \end{aligned} \tag{12}$$

On note ici que l'introduction de cette contrainte rend la contrainte (11) redondante. En effet, par contraposée de la contrainte (8), un processus non activable ne peut pas être activé. Néanmoins, la contrainte (11) est conservée afin de renforcer le pouvoir de propagation du solveur et de permettre une mise à jour plus efficace des horloges des processus non activables. De plus, il est important de noter qu'en cas d'interblocage, même si l'exécution courante doit s'arrêter, les processus ne sont pas activés et conservent donc leurs valeurs jusqu'à l'instant $t_f - 1$ grâce à la contrainte (12) (ou (11)), ce qui permettra de détecter les phénomènes de convergence et de divergence décrits dans la section 3.2.

(iii) Processus activables et activés. Un processus est *activable et activé* à l'étape t lorsqu'il est autorisé à modifier sa valeur locale et qu'il est effectivement sélectionné par le démon pour exécuter une transition à cette étape. Dans ce cas, le processus met nécessairement à jour son état conformément aux règles de l'algorithme de l'unisson asynchrone décrites dans l'algorithme 1. Ainsi, chaque processus activé doit mettre à jour sa valeur selon son statut d'autorisation : s'il est autorisé à l'incrémentation, il doit incrémenter sa valeur modulo K à la prochaine configuration ; et, si il est autorisé à réinitialiser sa valeur, elle doit devenir nulle à la prochaine configuration. Ces contraintes, décrites formellement ci-dessous, génèrent un nombre de clauses en $O(n * t_f * K)$:

$$\begin{aligned} clk_{p,t,v} \wedge inc_{p,t} \wedge act_{p,t} \rightarrow clk_{p,t+1,(v+1) \bmod K} \\ \forall p \in V, \forall t \in T \setminus \{t_f - 1\},\ \forall v \in M \end{aligned} \tag{13}$$

$$\begin{aligned} clk_{p,t,v} \wedge res_{p,t} \wedge act_{p,t} \rightarrow clk_{p,t+1,0} \\ \forall p \in V, \forall t \in T \setminus \{t_f - 1\},\ \forall v \in M \end{aligned} \tag{14}$$

Ainsi, l'ensemble des contraintes définies ci-dessus permet de modéliser de manière exhaustive toutes les exécutions valides de l'algorithme asynchrone. L'utilisation conjointe des variables d'activation et d'autorisation garantit une représentation compacte et efficace des transitions. Dans la section suivante, nous nous intéresserons à la formalisation des contraintes permettant d'analyser les propriétés d'autostabilisation de l'algorithme.

3.2 Analyse de l'autostabilisation

Cette section est consacrée à l'analyse du comportement d'autostabilisation de l'algorithme de l'unisson asynchrone. Pour ce faire, nous distinguons deux aspects complémentaires : d'une part, la convergence, qui garantit l'atteinte d'une configuration légitime à partir de toute configuration initiale en un nombre fini d'itérations ; d'autre part, la divergence, qui permet d'identifier et de caractériser les exécutions qui ne peuvent pas mener à un état légitime. Cette analyse doit prendre en compte l'ensemble des exécutions possibles induites par l'asynchronisme de l'algorithme.

3.2.1 Convergence

Nous rappelons que la convergence est établie si, à partir de toute configuration initiale, l'exécution asynchrone de l'algorithme aboutit à une configuration légitime dans laquelle la valeur d'horloge de chaque processus p est en retard d'au plus un incrément par rapport à chacune des horloges de ses voisins. Plus précisément, si p possède la valeur d'horloge $v \in M$, alors la valeur d'horloge u d'un voisin $q \in N(p)$ appartient nécessairement à l'ensemble $\{v-1 \bmod K,\ v,\ v+1 \bmod K\}$. On rappelle ici que, les horloges étant modulaires, l'écart entre $K-1$ et 0 n'est que d'un seul incrément. De plus, l'ensemble des configurations légitimes est clos : toute configuration accessible depuis une configuration légitime est également légitime.

Plutôt que de considérer toutes les exécutions asynchrones directement, nous raisonnons par contradiction. Nous chercherons à vérifier s'il existe un chemin d'exécution asynchrone issu d'une configuration initiale arbitraire où l'algorithme ne converge pas, c'est-à-dire où la configuration à l'étape $t_f - 1$ comporte au moins un processus p possédant la valeur v et un voisin $q \in N(p)$ avec une valeur u dont l'écart par rapport à v est supérieur à un incrément, c'est-à-dire que $u \notin \{v-1 \bmod K,\ v,\ v+1 \bmod K\}$. Cette situation viole la propriété de convergence sur l'écart légitime et il suffit donc de considérer un seul chemin qui échoue pour aboutir à une contradiction ; sinon, tous les chemins d'exécution asynchrone possibles respectent la propriété de convergence et l'algorithme converge au plus tard après t_f configurations.

Pour formaliser cette propriété, on introduit les **variables de retard légitime** $obhd_{p,q,v}$ pour deux processus distincts $(p,q) \in V^2$ et $v \in M$, qui seront affecté à $vrai$ si le processus p possède une valeur d'horloge avec un écart d'au plus 1 par rapport à q à la dernière configuration $t_f - 1$. Ainsi, la non-convergence est clairement exprimée par l'existence d'un couple de processus voisins (p,q) et d'une valeur d'horloge $v \in M$ du processus p à la configuration $t_f - 1$ tel que q n'est pas en retard légitime par rapport à p ($obhd_{p,q,v} = faux$). Pour faciliter la réécriture sous format FNC, nous introduisons également les **variables d'horloges légitimes** $oclk_{p,v}$ pour $p \in V$ et $v \in M$, qui seront affecté à $vrai$ si le processus p possède une valeur d'horloge avec un écart d'au plus 1 par rapport à la valeur v à la dernière configuration $t_f - 1$. Ces contraintes, permettant l'analyse de la converge, sont présentées ci-dessous, générant un nombre de clauses borné en $O(n^2 * K)$:

$$\bigvee_{\substack{p \in V \\ q \in N(p) \\ v \in M}} \overline{obhd_{p,q,v}} \tag{15}$$

$$obhd_{p,q,v} \leftrightarrow (\overline{clk_{p,t_f-1,v}} \ \vee \ oclk_{q,v}) \quad \forall p \in V,\ \forall q \in V \setminus \{p\},\ \forall v \in M \tag{16}$$

$$oclk_{p,v} \leftrightarrow (clk_{p,t_f-1,v} \ \vee \ clk_{p,t_f-1,(v-1) \bmod K} \ \vee \ clk_{p,t_f-1,(v+1) \bmod K}) \quad \forall p \in V,\ \forall v \in M \tag{17}$$

Ainsi, l'algorithme de l'unisson asynchrone converge sur un graphe connexe non orienté de taille n en t_f configurations si et seulement si la formule induite par les contraintes présentées est insatisfiable. Cette formule permettant d'analyser la convergence contient globalement $O(n^2 * t_f * K^2)$ clauses.

Théorème 1. *L'algorithme de l'unisson asynchrone converge sur un graphe connexe non orienté de taille n en au plus t_f configurations ssi la formule induite par les contraintes (1-17) est insatisfiable.*

3.2.2 Divergence

La divergence de l'algorithme de l'unisson asynchrone correspond à l'existence d'une exécution qui n'atteint jamais une configuration légitime. Autrement dit, cela implique nécessairement l'existence d'un interblocage ou d'un cycle de configurations illégitimes (puisque l'ensemble des configurations légitimes est clos). En raison de l'asynchronisme, chaque configuration peut engendrer plusieurs exécutions possibles. Ainsi, pour établir la divergence, il suffit de montrer l'existence d'un interblocage ou d'un cycle de configurations identiques et illégitimes le long d'un même chemin d'exécution asynchrone.

Plus formellement, dans le cas d'un cycle, on recherche deux configurations $0 \leq t_1 < t_2 < t_f$ formant une paire de configurations identiques et illégitimes. Sans perte de généralité, on se restreint au cas où $t_1 = 0$, c'est-à-dire à la comparaison entre une configuration atteinte à l'étape $t \neq 0$ et la configuration initiale. Ce choix est justifié par le fait que toute configuration peut être considérée comme initiale, et donc l'analyse peut toujours être ramenée à cette configuration. Afin de garantir que le cycle considéré ne contient que des configurations illégitimes, il suffit d'imposer que la configuration initiale soit elle-même illégitime. Comme annoncé précédemment, les cas d'interblocage sont également

détecté à l'aide de cette méthode car, dans notre formalisation, ils deviennent des cas particuliers de cycles. En effet, toute exécution atteignant une configuration interbloquée peut aussi être ramené à une exécution où cette configuration illégitime est initiale. Dans ce cas, grâce à la conservation des valeurs d'horloge réalisée dans notre formalisation, un cycle sera détecté en une seule itération puisque les deux premières configurations seront identiques.

Pour raisonner sur l'existence de cycles, on introduit les **variables de cycles** cyc_t, pour $t \in T$, indiquant l'existence d'un cycle entre la configuration initiale et la configuration t. L'existence d'un cycle de configurations illégitimes est alors modélisée par les contraintes (18) et (19), générant deux clauses. La première clause exprime l'existence d'au moins une configuration non initiale formant un cycle tandis que la deuxième impose que la configuration initiale soit illégitime (et donc celle à t aussi), en garantissant l'existence d'un retard non légitime entre un processus et l'un de ses voisins.

$$\bigvee_{t \in T \setminus \{0\}} cyc_t \tag{18}$$

$$\bigvee_{\substack{p \in V \\ q \in N(p) \\ v \in M}} \overline{obhd_{p,q,v}} \tag{19}$$

Il est important de noter que la contrainte (19) a une sémantique légèrement différente de celle de la convergence (cf. (15)) puisqu'on l'applique à la première configuration au lieu de la dernière. Il est donc nécessaire de mettre à jour la définition des variables de retard et d'écart légitimes afin prendre en compte cette nouvelle sémantique. De plus, afin de définir la sémantique des variables de cycle cyc_t, il est nécessaire d'imposer que, lorsque cyc_t est vraie, la configuration à l'étape t coïncide avec la configuration initiale. À cette fin, nous introduisons les **variables de similarité d'horloge** $sim_{p,t,v}$, indiquant que le processus p possède la valeur v simultanément aux configurations 0 et t. L'ensemble de ces contraintes sont définies ci-dessous, générant un nombre de clauses borné en $O(n^2 * t_f * K)$.

$$\begin{aligned} obhd_{p,q,v} \leftrightarrow (&\overline{clk_{p,0,v}} \ \vee \ oclk_{q,v}) \\ &\forall p \in V,\ \forall q \in V \setminus \{p\},\ \forall v \in M \end{aligned} \tag{20}$$

$$\begin{aligned} oclk_{p,v} \leftrightarrow (&clk_{p,0,v} \ \vee \ clk_{p,0,(v-1) \bmod K} \\ &\vee \ clk_{p,0,(v+1) \bmod K}) \\ &\forall p \in V,\ \forall v \in M \end{aligned} \tag{21}$$

$$\begin{aligned} cyc_t \leftrightarrow &\bigvee_{v \in M} sim_{p,t,v} \\ &\forall t \in T \setminus \{0\}, \forall p \in V \end{aligned} \tag{22}$$

$$\begin{aligned} sim_{p,t,v} \leftrightarrow &\ clk_{p,0,v} \wedge clk_{p,t,v} \\ &\forall t \in T \setminus \{0\},\ \forall p \in V,\ \forall v \in M \end{aligned} \tag{23}$$

Ainsi, la formule induite par les contraintes (1-14 ;18-23) est satisfiable si et seulement si l'algorithme de l'unisson asynchrone diverge en au plus t_f configurations. En effet, un modèle de cette formule permet d'exhiber une configuration initiale ainsi qu'une branche de l'arbre des exécutions asynchrone contenant un cycle de configurations illégitimes (y compris en cas d'interblocage). En termes de complexité, l'encodage global reste du même ordre que celui de la convergence, soit un nombre de clauses borné en $O(n^2 * t_f * K^2)$.

Théorème 2. *L'algorithme de l'unisson asynchrone diverge sur un graphe connexe non orienté de taille n en au plus t_f configurations ssi la formule induite par les contraintes (1-14 ;18-23) est satisfiable.*

4 Évaluation expérimentale

Nous avons considéré trois topologies classiques : anneaux, chaînes et étoiles. Pour chaque type de topologie, nous avons généré différentes instances en variant les tailles de graphes n de 3 jusqu'à 5 noeuds, et la période K de $n+1$ à n^2+1. Ainsi, on obtient 41 instances de convergence et divergence (CNV et DIV) correspondant respectivement aux formules induites par les contraintes (1-17) et (1-14 ;18-23) pour chaque topologie, et totalisant 246 instances. Nous avons utilisé l'encodage Cardinality Network [3] pour réécrire la contrainte de cardinalité (1)[3]. Enfin, on fixe $t_f = 100$ pour toutes les instances. Nos modèles ont été encodés en Python en utilisant la bibliothèque PySAT[4] [17]. Nous avons résolu ces instances en utilisant le solveur Cadical [6]. Les tests ont été réalisés sur la plateforme MatriCS[5], sur une machine équipée d'un processeur Intel Core i7 cadencé à 3,80 GHz, sous Ubuntu 22.04. Un temps limite de 3600 s a été alloué à chaque instance.

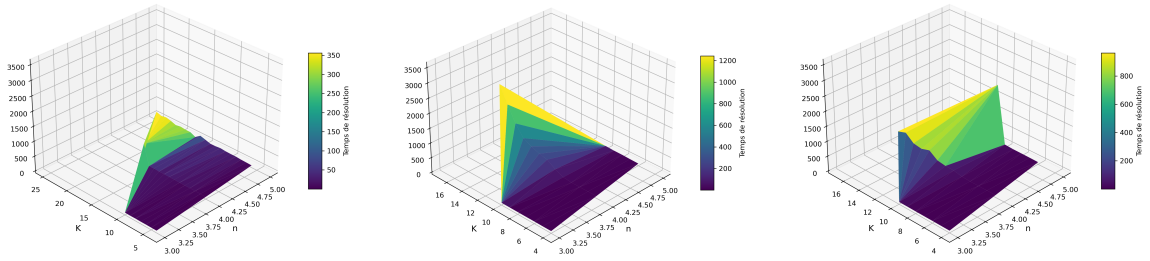
Tout d'abord, nous analysons le temps de résolution et le nombre d'instances résolues sur différentes topologies par taille de graphe. Les résultats obtenus sont présentés dans le tableau 3, en fonction de la taille du graphe et de la propriété analysée (convergence/divergence). Par ailleurs, les figures 4 et 5 illustrent l'évolution du temps de résolution des instances générées respectivement pour la vérification de la convergence et de la divergence, en fonction de la taille du graphe n et de la période K sur les différentes topologies. Il est remarquable que le taux d'évolution du temps de résolution dans les anneaux est beaucoup plus faible par rapport aux chaînes et aux étoiles.

Les résultats indiquent que, sur l'ensemble des instances générées, le pourcentages d'instances résolues pour la détection de la convergence (resp. divergence) est de $97,56\%$ (resp. $97,56\%$) pour les anneaux, $60,98\%$ (resp. $85,37\%$) pour les chaînes et $63,41\%$ (resp. 100%) pour les étoiles. Plus précisément, pour les tailles $n = 3$ et $n = 4$, toutes les instances (c'est-à-dire, 7 pour $n = 3$ et 13 pour $n = 5$) ont

3. Pour encoder la contrainte $\sum_{i=1}^{h} l_i = 1$, cet encodage génère $O(h)$ clauses et variables auxiliaires.
4. https://pysathq.github.io/
5. https://www.matrics.u-picardie.fr/catalogue-de-services/compte-matrics/

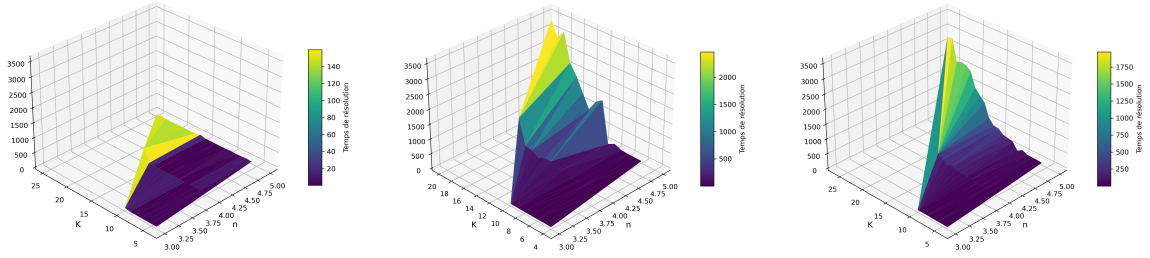
n	$\#I$	**Anneaux**		**Chaînes**		**Étoiles**	
		CNV	**DIV**	**CNV**	**DIV**	**CNV**	**DIV**
3	7	11.08 [7]	36.87 [7]	14.25 [7]	25.42 [7]	12.49 [7]	26.06 [7]
4	13	871.83 [13]	438.01 [13]	6947.44 [13]	2613.05 [13]	2160.56 [13]	1158.38 [13]
5	21	1584.10 [20]	170.98 [20]	102.51 [5]	18589.88 [15]	2020.86 [6]	20362.94 [21]
Total	**41**	**2467.01 [40]**	**645.86 [40]**	**7064.20 [25]**	**21228.35 [35]**	**4193.90 [26]**	**21547.38 [41]**

TABLE 3 – Temps de résolution en secondes pour les topologies en anneau, chaîne et étoile. Pour chaque topologie, les sommes des temps de résolution pour la détection de convergence (CNV) et divergence (DIV) sur toutes les périodes (K) pour une taille de graphe donnée (n) sont reportées. Pour une taille n donnée, la colonne $\#I$ indique le nombre d'instances générées avec CNV ou DIV et le nombre d'instances résolues est indiqué entre '[]'.



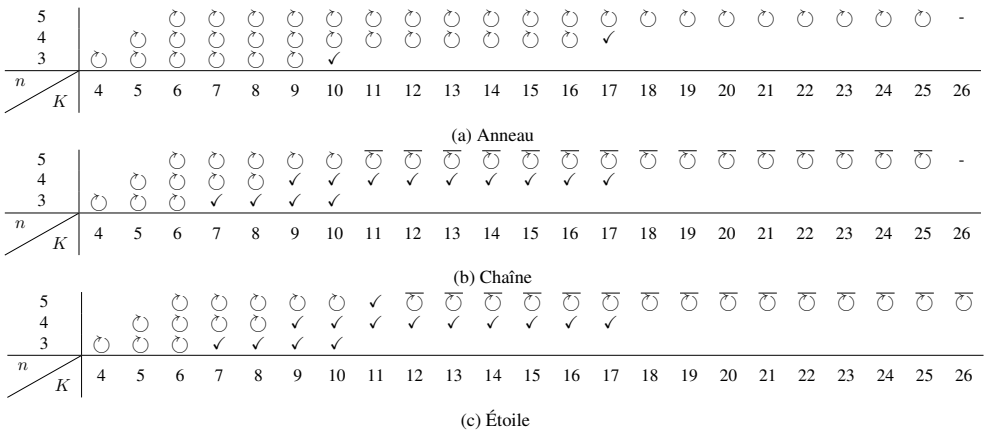
(a) Anneau (b) Chaîne (c) Étoile

FIGURE 4 – Temps de résolution en secondes pour la convergence en fonction du nombre de processus n et de la période K.



(a) Anneau (b) Chaîne (c) Étoile

FIGURE 5 – Temps de résolution en secondes pour la divergence en fonction du nombre de processus n et de la période K.



(a) Anneau

(b) Chaîne

(c) Étoile

FIGURE 6 – Résultats de convergence et de divergence de l'algorithme de l'unisson asynchrone sur différents types de graphes pour des tailles $n = 3, 4, 5$ et périodes $K = n + 1 \dots n^2 + 1$.

été résolues à la fois pour la convergence et la divergence pour toutes les topologies. Avec les graphes de plus grande taille $n = 5$, 20 instances de convergence et de divergence relatives à la topologie en anneau ont été résolues montrant un comportement divergent. Avec les topologies en chaîne et en étoile, nous remarquons que la résolution est beaucoup plus difficile surtout dans le cas de la divergence avec seulement 5 instances dans les chaînes (respectivement 6 dans les étoiles) pour le cas de la convergence menant à une baisse drastique du nombre d'instances résolues. En effet, seules 5 ou 6 instances sont résolues sur les 21 instances générées. Par ailleurs, nous remarquons que toutes les instances ont été résolues dans le cas de la divergence pour les étoiles. Par contre, seulement 15 instances instances ont été résolues dans le cas de la divergence pour les chaînes.

En terme de temps de résolution, pour un même nombre d'instance résolues sur les tailles de graphe de 3 et 4, nous remarquons que le modèle est plus performant sur les graphes en anneaux que sur les chaînes et les étoiles avec une accélération de $87,32\%$ et de $82,00\%$ par rapport aux graphes en chaînes et de $59,37\%$ et $59,91\%$ par rapport aux graphes en étoiles respectivement dans le cas de la convergence et de la divergence. Pour les graphes de taille 5, le modèle a résolu dans le cas des anneaux 20 instances de plus entre convergence et divergence par rapport aux chaînes et 13 instances de plus par rapport aux étoiles.

Les résultats présentés dans la Figure 6 illustrent les comportements exhibés de l'algorithme d'unisson asynchrone en termes de convergence et de divergence sur les différents types de topologies (anneaux, chaînes et étoiles) en fonction de la taille n et de la période K. Les signes (✓), (↻), ($\overline{↻}$) et ($-$) indiquent respectivement que l'algorithme converge, diverge, ne diverge pas dans la limite du temps alloué ou que les instances de convergence et de divergence n'ont pas été résolues. Pour toutes les topologies, tous les comportements ont pu être exhibés pour les graphes de taille 3 et 4, révélant une convergence alignée avec les résultats prouvés dans [14] lorsque $K = n^2 + 1$. En revanche, pour les graphes de taille 5, certaines instances n'ont pas été résolues dans la limite d'une heure. Ceci est traduit par la complexité accrue du modèle dans des graphes de taille plus grande.

Nous remarquons une tendance claire de convergence et de divergence pour les topologies et tailles étudiées. Tout, d'abord, la convergence est constatée pour toutes les instances résolues avec $K = n^2 + 1$. De plus, on arrive à constater la divergence pour $K = n^2$ pour les anneaux de taille $n \in \{3, 4, 5\}$. Ces résultats s'alignent avec ceux établis dans [14] pour la convergence et montrent que la borne de $n^2 + 1$ est optimale dans le cadre général. On remarque également que la divergence est exhibée pour $K \leq 2n$ sur les topologies en chaîne et en étoile tandis que les instances qui ont été résolues au dessus de cette borne sont toujours convergentes. Ces observations nous conduisent à conjecturer que l'algorithme de l'unisson asynchrone est autostabilisant pour la borne raffinée $K \geq 2n+1$ pour les structures de graphes acycliques.

5 Conclusion

Dans cet article, nous avons présenté une approche formelle basée sur SAT pour analyser le comportement d'un algorithme autostabilisant d'unisson asynchrone. Plus précisément, notre étude porte sur la détection des cas où cet algorithme présente un comportement de convergence ou de divergence. En modélisant les états de l'algorithme ainsi que son exécution à l'aide de contraintes logiques, nous avons montré comment les solveurs SAT peuvent être exploités pour démontrer l'autostabilisation de l'unisson asynchrone appliqué à différentes topologies classiques et en tenant en compte de différentes périodes ou, à défaut, pour exhiber sa divergence. Nos résultats nous ont permis d'établir l'optimalité de la borne de convergence établie dans [14] ($K = n^2+1$). Cependant, nos observations laissent conjecturer que cette borne pourrait être raffinée sous certaines hypothèses sur la structure topologique des réseaux considérés. En particulier, l'acyclicité pourrait permettre d'obtenir une borne de convergence plus précise, notamment pour des périodes à partir de $K = 2n + 1$.

Cette étude ouvre la voie à plusieurs pistes de recherche. Tout d'abord, l'extension de notre analyse à d'autres types de graphes classiques, tels que les grilles, les tores ou bien des graphes aléatoires, permettrait d'analyser de façon plus approfondie l'impact d'autres topologies distribuées sur l'autostabilisation de l'algorithme. Il serait également intéressant d'étudier d'autres formulations plus compactes des contraintes afin de réduire la complexité du modèle. De surcroît, pour la contrainte de convergence, il serait intéressant d'envisager une analyse plus fine en établissant des contraintes redondantes sur l'ensemble des configurations au lieu de se focaliser uniquement sur la vérification de la synchronisation au niveau de la dernière à l'instar de [18]. Enfin, il serait pertinent de prouver formellement l'exactitude de la borne établie ou encore de la raffiner davantage. Ces perspectives favorisent une compréhension plus approfondie des dynamiques de synchronisation dans les systèmes distribués asynchrones complexes, et à l'élaboration de nouvelles méthodes plus optimisées pour couvrir un spectre plus vaste d'exemples d'applications.

Remerciements

Ce travail est partiellement soutenu par les projets ANR-24-CE23-6126 (BforSAT) et ANR-22-EXES-0009 (PIA4 MAIA, France 2030), financés par l'agence nationale de recherche (ANR). Il a bénéficié d'un accès aux ressources HPC de la « Plateforme MatriCS » de l'Université de Picardie Jules Verne, cofinancée par l'Union Européenne avec le Fonds Européen de Développement Régional (FEDER) et le Conseil Régional des Hauts-De-France.

Références

[1] Karine Altisen, Stéphane Devismes, and Erwan Jahier. sasa : a simulator of self-stabilizing algorithms. *Comput. J.*, 66(4) :796–814, 2023.

[2] Anish Arora, Shlomi Dolev, and Mohamed G. Gouda. Maintaining digital clocks in step. *Parallel Processing Letters*, 1 :11–18, 1991.

[3] Roberto Asín, Robert Nieuwenhuis, Albert Oliveras, and Enric Rodríguez-Carbonell. Cardinality networks and their applications. In Oliver Kullmann, editor, *Theory and Applications of Satisfiability Testing - SAT 2009, 12th International Conference, SAT 2009, Swansea, UK, June 30 - July 3, 2009. Proceedings*, volume 5584 of *Lecture Notes in Computer Science*, pages 167–180. Springer, 2009.

[4] Armin Biere. Handbook of satisfiability. In *Frontiers in Artificial Intelligence and Applications*, pages 75–98. IOS Press, 2009.

[5] Armin Biere et al. Symbolic model checking without bdds. In *Proceedings of the International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, pages 193–207, 1999.

[6] Armin Biere, Tobias Faller, Katalin Fazekas, Mathias Fleury, Nils Froleyks, and Florian Pollitt. CaDiCaL 2.0. In Arie Gurfinkel and Vijay Ganesh, editors, *International Conference on Computer Aided Verification, CAV*, volume 14681 of *LNCS*, pages 133–152. Springer, 2024.

[7] Armin Biere, Marijn Heule, Hans van Maaren, and Toby Walsh, editors. *Handbook of Satisfiability - Second Edition*, volume 336 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2021.

[8] L. Blin, C. Johnen, G. Le Bouder, and F. Petit. Silent anonymous snap-stabilizing termination detection. In *41st International Symposium on Reliable Distributed Systems, (SRDS'22)*, pages 156–165, 2022.

[9] C. Boulinier and F. Petit. Self-stabilizing wavelets and rho-hops coordination. In *22nd IEEE International Symposium on Parallel and Distributed Processing, (IPDPS'08)*, pages 1–8, 2008.

[10] C. Boulinier, F. Petit, and V. Villain. When graph theory helps self-stabilization. In *23rd Annual Symposium on Principles of Distributed Computing, (PODC'04)*, pages 150–159, 2004.

[11] E. M. Clarke and E. A. Emerson. Design and synthesis of synchronization skeletons using branching-time temporal logic. In *Proceedings of the Workshop on Logic of Programs*, volume 131 of *LNCS*, pages 52–71. Springer, 1982.

[12] Stephen A. Cook. The complexity of theorem-proving procedures. In Michael A. Harrison, Ranan B. Banerji, and Jeffrey D. Ullman, editors, *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing, May 3-5, 1971, Shaker Heights, Ohio, USA*, pages 151–158. ACM, 1971.

[13] Jean-Michel Couvreur, Nissim Francez, and Mohamed G. Gouda. Asynchronous unison (extended abstract). In *Proceedings of the 12th International Conference on Distributed Computing Systems, Yokohama, Japan, June 9-12, 1992*, pages 486–493. IEEE Computer Society, 1992.

[14] Jean-Michel Couvreur, Nissim Francez, and Mohamed G. Gouda. Asynchronous unison (extended abstract). In *The 12th International Conference on Distributed Computing Systems (ICDCS)*, pages 486–493. IEEE Computer Society, 1992.

[15] A. K. Datta, S. Devismes, and L. L. Larmore. A silent self-stabilizing algorithm for the generalized minimal *k*-dominating set problem. *Theoretical Computer Science*, 753 :35–63, 2019.

[16] Edsger W. Dijkstra. Self-stabilizing systems in spite of distributed control. *Commun. ACM*, 17(11) :643–644, 1974.

[17] Alexey Ignatiev, António Morgado, and João Marques-Silva. Pysat : A python toolkit for prototyping with SAT oracles. In Olaf Beyersdorff and Christoph M. Wintersteiger, editors, *Theory and Applications of Satisfiability Testing - SAT 2018 - 21st International Conference, SAT 2018, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 9-12, 2018, Proceedings*, volume 10929 of *Lecture Notes in Computer Science*, pages 428–437. Springer, 2018.

[18] Asma Khoualdia, Sami Cherif, Stéphane Devismes, and Léo Robert. Analyzing self-stabilization of synchronous unison via propositional satisfiability. In Maria Garcia de la Banda, editor, *31st International Conference on Principles and Practice of Constraint Programming, CP 2025, Glasgow, Scotland, August 10-15, 2025*, volume 340 of *LIPIcs*, pages 19 :1–19 :21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2025.

[19] João P. Marques-Silva and Karem A. Sakallah. Grasp : A search algorithm for propositional satisfiability. *IEEE Transactions on Computers*, 48(5) :506–521, 1999.

[20] Olivier Roussel and Vasco Manquinho. Pseudo-boolean and cardinality constraints. In Armin Biere, Marijn Heule, Hans van Maaren, and Toby Walsh, editors, *Handbook of Satisfiability - Second Edition*, volume 336 of *Frontiers in Artificial Intelligence and Applications*, pages 1087–1129. IOS Press, 2021.

Extraction des politiques pour la programmation par ensembles réponses quantifiée

Martín Diéguez[1], Igor Stéphan[1]

[1] Université d'Angers, LERIA

martin.dieguezlodeiro | igor.stephan@univ-angers.fr

Résumé

La programmation par ensembles réponses quantifiée (Quantified Answer Set Programming, QASP) étend la programmation par ensembles réponses (Answer Set Programming, ASP) en autorisant la quantification des variables propositionnelles, à l'instar des QBF (Quantified Boolean Formulas) vis-à-vis de la logique propositionnelle. Dans cet article, nous interprétons les modèles de formules QASP en termes de politiques qui représentent des stratégies qui déterminent comme les variables existentiellement quantifiées doivent être assignées selon les conditions fixées par les variables universellement quantifiées. La contribution principale est un algorithme, qui extrait les politiques selon la sémantique QASP, inspiré par la sémantique de la logique de l'équilibre pour des théories ASP générales.

Mots-clés

QASP, quantification, ASP.

Abstract

Quantified Answer Set Programming (QASP) extends Answer Set Programming (ASP) by allowing quantification over propositional variables, similar to Quantified Boolean Formulas (QBF). In this paper, we interpret models of QASP formulas in terms of policies*, which represent decision-making strategies that determine how existentially quantified variables should be assigned, given the conditions set by universally quantified variables. As a main contribution, we present an algorithm for policy extraction under QASP semantics, inspired by the Equilibrium Logic semantics for general ASP theories.*

Keywords

QASP, quantification, ASP.

Ce que vous lirez en anglais

Ayant ses racines dans la programmation logique, la programmation par ensembles réponses (*Answer Set Programming* [5], ASP) est un formalisme pour le raisonnement non monotone (*nonmonotonic reasoning*, NMR) et la représentation de connaissances (*Knowledge Representation*, KR) qui résoult des problèmes complexes impliquant de la recherche combinatoire, de l'optimisation et du raisonnement dans l'incertain. En ASP, un problème est encodé en un ensemble de règles et les modèles (ensembles réponses) représentent les solutions. En marge de la traditionnelle sémantique des modèles stables, il existe des interprétations alternatives des programmes logiques selon la sémantique des ensembles réponses [7] incluant celle de la logique de l'équilibre (EL) de Pearce [8] qui utilise une base monotone et une condition de minimalité pour induire la non monotonie. Pour des théories propositionnelles, le problème de satisfaction en ASP se situe au second niveau de la hiérarchie polynomiale (Σ_2^p) [9]. Les approches telles que la sémantique stable-instable [4] étendent ASP pour résoudre les problèmes au-delà de cette classe de complexité en intégrant des programmes logiques comme des oracles. Une autre approche est le système ASP(Q) [1] [1], qui étend la syntaxe ASP avec des quantifications existentielles et universelles portant sur les modèles stables du programme plutôt que sur les atomes.

Dans le cas classique, les formules booléennes quantifiées (*Quantified Boolean Formulas* [3, Chapter 31], QBFs) étendent la logique propositionnelle avec des quantifications du second ordre autorisant l'expression de problèmes au-delà de la classe de complexité Σ_2^p, étant donné que le problème de satisfiabilité des QBF se situe dans PSPACE-complet [11]. La programmation par ensembles réponses quantifiée (QASP) [10, 6], similairement étend ASP en incorporant des quantificateurs propositionnels de manière analogue à QBF. Comme en QBF, les quantifications existentielles et universelles s'appliquent sur la valeur de vérité de l'atome propositionnel quantifié. La différence clé réside dans la sémantique : en QASP, la valeur de vérité assignée à l'atome doit être consistante avec les modèles stables du programme. Il existe deux sémantiques principales pour les programmes QASP. Dans la sémantique proposée par Fandinno et al. [6], un atome p est forcé a être vrai dans le modèle stable en ajoutant la formule $\neg\neg p$ au contexte. En revanche, dans la sémantique définie dans Stephan [10], p est rendu vrai simplement en ajoutant p au contexte. Les deux sémantiques sont identiques lorsqu'il s'agit de déterminer si un atome est faux dans un modèle stable. En termes d'implantation, la sémantique de [6] est réalisée en un outil appelé `qasp2qbf` [2], qui a été utilisé pour résoudre des

1. https://www.mat.unical.it/ricca/downloads/qasp-0.1.2.jar
2. https://github.com/potassco/qasp2qbf.

problèmes de planification sous incertitude.

En QASP, les solutions à un programme logique quantifié peuvent être interprétées comme des stratégies dans un jeu à deux joueurs avec information complète. Le premier joueur tente de construire les modèles stables du programme ASP, et le programme est considéré comme une instance positive du problème QASP si et seulement si une stratégie gagnante pour le premier joueur. L'outil `qasp2qbf` réduit un programme QASP à une QBF, dont la satisfiabilité est décidée par un solveur QBF. Si la QBF est satisfiable, le solveur fournit un assignement pour le bloc initial de quantificateurs existentiels. Cela signifie que si le programme QASP commence avec un quantificateur universel, le solveur ne peut calculer que le caractère satisfiable ou non, sans produire une stratégie. En revanche, l'approche présentée dans [10] remplace les quantificateurs existentiels avec des fonctions de Skolem explicitant les choix du joueur existentiel. Cette méthode est donc en mesure de calculer toutes les stratégies qui satisfont le programmme QASP.

Dans le présent article, nous explorons les sémantiques de [6] et de [10] sous la perspective de la logique de l'équilibre. Nous introduisons la restriction suivante : nous ne considérons pas un langage logique spécifique mais, à la place, nous traduisons chaque règle d'un programme logique $(H \leftarrow B)$ en une implication $(B \rightarrow H)$ en logique propositionnelle. La négation par défaut (`not`) est remplacée par la négation ($\neg$) de la logique propositionnelle, et les formules propositionnelles peuvent être imbriquées de manière arbitraire.

La première contribution de cet article démontre que les deux sémantiques ne sont pas équivalentes. La deuxième contribution est l'extension de la notion de politique du cas classique à la logique propositionnelle de Gödel quantifiée QG_3 [2], qui fournit une base monotone pour sélectionner les politiques minimales. La troisième contribution est un algorithme qui prend une QBF en entrée et calcule l'ensemble des politiques selon la sémantique de [6]. Cet algorithme combine les politiques QBF et QG_3 durant l'exécution.

Références

[1] G. Amendola, F. Ricca, and M. Truszczynski. Beyond NP : Quantifying over answer sets. *Theory and Practice of Logic Programming*, 19(5-6) :705–721, 2019.

[2] M. Baaz, A. Ciabattoni, and R. Zach. Quantified propositional Gödel logics. In *Proceedings of the 7th International Conference on Logic for Programming and Automated Reasoning (LPAR)*, volume 1955 of *LNCS*, pages 240–256. Springer, 2000.

[3] A. Biere, M. Heule, H. van Maaren, and T. Walsh, editors. *Handbook of Satisfiability*, volume 336 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2021.

[4] B. Bogaerts, T. Janhunen, and S. Tasharrofi. Stable-unstable semantics : Beyond NP with normal logic programs. *Theory and Practice of Logic Programming*, 16(5-6) :570–586, 2016.

[5] G. Brewka, T. Eiter, and M. Truszczyński. Answer set programming at a glance. *Communications of the ACM*, 54(12) :92–103, 2011.

[6] J. Fandinno, F. Laferriere, J. Romero, T. Schaub, and T. Son. Planning with incomplete information in quantified answer set programming. *Theory and Practice of Logic Programming*, 21(5) :663–679, 2021.

[7] V. Lifschitz. Thirteen definitions of a stable model. In *Fields of Logic and Computation : Essays Dedicated to Yuri Gurevich on the Occasion of His 70th Birthday*, pages 488–503. Springer, 2010.

[8] D. Pearce. Equilibrium logic. *Annals of Mathematics and Artificial Intelligence*, 47(1-2) :3–41, 2006.

[9] D. Pearce, H. Tompits, and S. Woltran. Encodings for equilibrium logic and logic programs with nested expressions. pages 306–320.

[10] I. Stéphan. QASP ou la programmation par ensembles réponses quantifiée. In *Actes du dix-neuvième congrès national sur la Reconnaissance de Formes et l'Intelligence Artificielle (RFIA)*, 2014. In French.

[11] L.J. Stockmeyer and A.R. Meyer. Word problems requiring exponential time. In *Proceedings of the 5th annual ACM symposium on Theory of computing (STOC)*, pages 1–9, 1973.

Vers un système de raisonnement neurosymbolique ancré dans des représentations schématiques

François Olivier[1], Zied Bouraoui[2]

[1]STIH, Sorbonne Université, Paris, France
[2]CRIL, CNRS & Université d'Artois, Lens, France

francois.olivier@sorbonne-universite.fr bouraoui@cril.fr

Résumé

Cet article présente Embodied-LM*, un prototype de système neurosymbolique qui ancre le raisonnement dans des représentations schématiques inspirées de la théorie cognitive des image schemas. Le modèle de langage identifie les structures schématiques pertinentes dans des énoncés et les traduit en programmes logiques exécutés par un système de raisonnement ASP enrichi de prédicats spatiaux. Des expériences sur des tâches de déduction logique montrent que ces représentations permettent un raisonnement interprétable et efficace.*

Mots-clés

IA neurosymbolique, représentation des connaissances, linguistique cognitive, raisonnement spatial, modèles de langage

Abstract

This paper presents Embodied-LM*, a prototype neurosymbolic system that grounds reasoning in schematic representations inspired by the cognitive theory of image schemas. The language model identifies relevant schematic structures in sentences and translates them into logical programs executed by an ASP-based reasoning system enriched with spatial predicates. Experiments on logical deduction tasks show that these representations enable interpretable and effective reasoning.*

Keywords

Neurosymbolic AI, knowledge representation, cognitive linguistics, spatial reasoning, language models

1 Le raisonnement incarné

Le raisonnement à partir d'énoncés en langage naturel pose un problème central en intelligence artificielle : comment représenter de manière structurée les contraintes décrites dans le langage afin de permettre une inférence fiable ? Si les systèmes récents fondés sur des modèles de langage atteignent des performances impressionnantes, leurs processus de raisonnement restent souvent difficiles à inspecter et peuvent produire des inférences erronées, tout en demeurant peu alignés avec les mécanismes cognitifs humains [13, 11, 23]. Une approche consiste alors à compléter ces systèmes par un module de raisonnement symbolique qui ancre l'inférence dans des représentations structurées proches de celles mobilisées par l'humain [17, 16].

Dans le cadre de la théorie de la cognition incarnée, Johnson a introduit les *image schemas* [7], définis comme des structures récurrentes abstraites issues de nos interactions avec le monde physique, qui organisent aussi bien la pensée concrète qu'abstraite [12, 6]. Par exemple, le schéma CONTAINER, dérivé de nos expériences d'objets placés dans des récipients, peut être mobilisé pour comprendre des phrases comme "Bill est dans la maison", mais aussi des expressions plus abstraites comme "Berlin est en Allemagne", "she is in love" ou encore des expressions mathématiques comme $2 \in \mathbb{N}$ [9, 10]. Le schéma PATH, issu de nos expériences de déplacement d'un point de départ vers un point d'arrivée, structure de la même manière des expressions comme "ne pas s'arrêter en si bon chemin", mais aussi des représentations plus abstraites comme la ligne mentale des nombres [10] ou des comparaisons temporelles. Le schéma BALANCE, lié à l'expérience corporelle de l'équilibre, peut quant à lui intervenir dans la compréhension d'expressions comme "pencher en faveur de quelqu'un", ou encore dans la représentation des équations mathématiques comme deux termes devant rester en équilibre.

Lorsqu'on donne une réalisation graphique ou diagrammatique à ces structures schématiques, l'information organisée par les *image schemas* prend la forme d'une configuration spatiale manipulable [15]. Certaines relations peuvent alors devenir immédiatement explicites, alors qu'elles n'étaient pas directement données dans le texte. Un exemple simple est celui de l'utilisation du schéma PATH. Considérons l'énoncé suivant : « Alice est plus âgée que Bill, et Charles est plus jeune que Bill ». Comprendre cet énoncé revient à projeter les individus sur une dimension ordonnée représentant l'âge, ce qui conduit à une configuration comme celle de la figure 1.

Une fois cette configuration construite, certaines conséquences deviennent immédiatement visibles : par exemple, Alice apparaît directement comme plus âgée que Charles, bien que cette relation ne soit pas explicitement énoncée. Inférer cette nouvelle information repose alors sur la simple observation de la structure même de la représentation, plutôt que sur une manipulation symbolique [21].

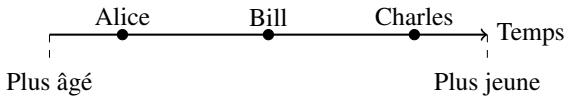


FIGURE 1 – Projection spatiale d'un ordre temporel au moyen du schéma PATH.

C'est précisément ce principe inférentiel, proche du raisonnement diagrammatique, que le système Embodied-LM cherche à formaliser dans un cadre de raisonnement automatique [17]. La section suivante en présente l'implémentation.

2 Implémentation

Le prototype Embodied-LM [17] repose sur un pipeline neurosymbolique simple exposé à la figure 2. À partir d'un énoncé en langage naturel, un modèle de langage produit un programme d'entrée où les contraintes logiques de l'énoncé sont interprétées métaphoriquement à l'aide de schémas spatiaux. Ce programme est ensuite résolu par un système en *Answer Set Programming* [8], enrichi de contraintes spatiales exprimées sous forme de contraintes polynomiales [1]. Les détails de la génération des différentes parties du programme d'entrée, ainsi que les aspects techniques du raisonneur symbolique sont décrits dans ce qui suit.

Module linguistique. Ce premier module utilise GPT-4 [19] par *prompting* pour interpréter l'énoncé au moyen d'*image schemas* et produire un programme logique décrivant le contexte du problème. Les entités mentionnées dans l'énoncé et les relations qui les relient sont ainsi identifiées puis traduites dans le vocabulaire spatial spécifié dans le prompt.

Extrait du prompt fourni à GPT-4

```
You are an expert at writing declarative
programs that encode the interpretation
of natural language narratives through image
schemas - recurring patterns derived from
sensorimotor experience that structure
human cognition. Declarative programs are
written in Answer Set Programming augmented
with spatial theory predicates.

OBJECT DECLARATION:
Objects in narratives correspond to geometric
objects in the program. Specify the geometric
sort of one or more objects by
"point(obj1;obj2;...)." ...

BUILTIN SPATIAL PREDICATES AVAILABLE:
- left(P1,P2): Point P1 is left to point P2.
- on(P1,L2): Point P1 is on line L2.
- ...

Generate the Clingo program with the
qualitative spatial predicates available to
encode the spatial schematic interpretation
of this natural language input.

{CONTEXT}
```

Dans le cas de l'énoncé introduit dans la section précédente, les individus sont ainsi représentés comme des points situés sur un segment abstrait correspondant à une dimension ordonnée. Le programme généré par ce module linguistique pourrait alors être le suivant.

Listing 1 – Programme généré pour l'énoncé "Alice est plus âgée que Bill, et Charles est plus jeune que Bill".

```
point(alice; bill; charles).
segment(timeline).

on(X, timeline) :- point(X).
left(alice, bill).
right(charles, bill).
```

GPT-4 n'a donc pas à inventer la signification des prédicats utilisés : il sélectionne, parmi un vocabulaire spatial fixé à l'avance, les prédicats qui correspondent aux contraintes exprimées dans l'énoncé. Une première utilisation du module symbolique vérifie que le programme généré est conforme aux spécifications données et relance la génération d'un programme si ce n'est pas le cas.
Lorsque l'énoncé est accompagné d'une question, le module linguistique est aussi chargé, par le biais d'un second prompt, d'étendre le programme par des règles de la forme `answer(a) :- condition`, représentant les différentes options de réponse. Ces règles sont ensuite ajoutées au programme initial.

Module de raisonnement. Le programme obtenu est ensuite exécuté par *clingo*, dont la syntaxe d'entrée est enrichie par des prédicats spatiaux [8, 3]. La sémantique de ces prédicats est définie analytiquement par des équations polynomiales portant sur les paramètres géométriques des objets, comme décrit dans le *Declarative Spatial Reasoning framework* [1, 24]. Par exemple, une relation comme *left* entre deux points peut être interprétée par la contrainte $\mathit{left}(a, b) \equiv x_a < x_b$, où x_a et x_b désignent les coordonnées des points a et b sur l'axe des abscisses. De même, une relation comme *on* entre un point et un segment peut être définie par des contraintes exprimant que le point appartient à la droite portée par le segment et se situe entre ses deux extrémités.
Pour la recherche de modèles, *clingo* traite la partie logique du programme en déterminant quelles relations doivent être satisfaites à partir des règles déclarées, tandis que les contraintes paramétriques associées à ces relations sont résolues par le solveur SMT Z3 [5]. Un modèle est généré lorsque les contraintes logiques et spatiales du programme sont satisfaites simultanément. Lorsqu'un même objet apparaît dans plusieurs relations, les contraintes correspondantes portent sur les mêmes paramètres géométriques. Par exemple, si le programme contient à la fois *left(a,b)* et *right(c,b)*, Z3 cherche une même affectation des coordonnées de a, b et c qui satisfasse simultanément $x_a < x_b$ et $x_c > x_b$.
Une option de réponse, comme par exemple `answer(a)`, est considérée comme correcte lorsqu'elle apparaît dans tous les modèles générés par le programme. Les solutions

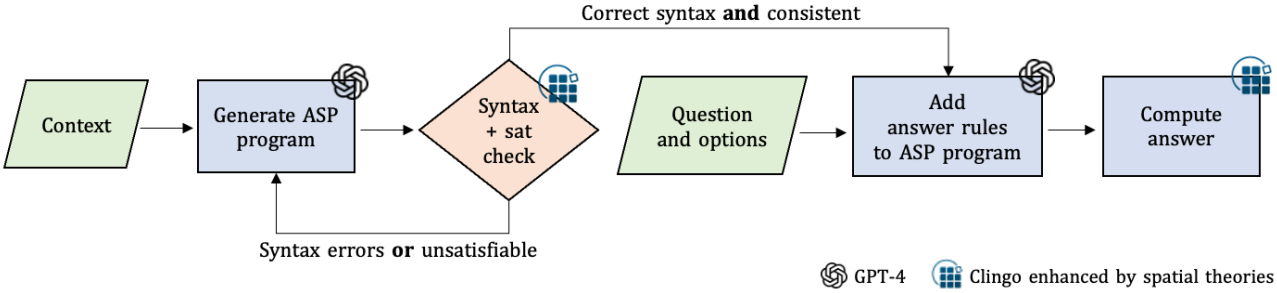


FIGURE 2 – Architecture d'Embodied-LM [17] : GPT-4 traduit l'énoncé en programme ASP enrichi par des prédicats spatiaux, puis le module de raisonnement combine *clingo* et Z3 afin de produire une réponse et, lorsque c'est utile, une représentation schématique du problème.

géométriques satisfaisant les contraintes spatiales constituent alors des *witnesses* qui attestent de la réalisabilité des relations spatiales impliquées dans le raisonnement. Ces *witnesses* peuvent également être visualisés [17] afin de produire des représentations schématiques du problème, analogues à celle de la figure 1, et ainsi rendre explicite la structure spatiale sur laquelle repose l'inférence.

3 Résultats et discussion

Nous évaluons Embodied-LM sur le dataset *LogicalDeduction* [22], composé de problèmes proches de l'exemple considéré ici : des entités sont ordonnées ou comparées à partir de relations exprimées en langage naturel, puis une question à choix multiple porte sur les conséquences logiques de ces prémisses.

System	Score
SymbCoT	93.00
Embodied-LM	**91.00**
VERUS-LM	88.67
Logic-LM	87.63
GPT4-CoT	75.25
GPT4	71.33

TABLE 1 – Performance sur le jeu de données *LogicalDeduction*. Résultats des autres systèmes repris de [4].

Comme le montre la Table 1, Embodied-LM atteint un score de 91%, proche des meilleurs systèmes récents. Nous avons également testé l'approche sur des *zebra puzzles* (aussi appelés *énigmes d'Einstein*), une famille classique de problèmes logiques consistant à déterminer une configuration d'objets satisfaisant un ensemble de contraintes relationnelles. Contrairement au dataset *LogicalDeduction*, ces problèmes sont plus naturellement représentés à l'aide du schéma CONTAINER, et Embodied-LM parvient également à les résoudre.

Ces expériences fournissent une validation conceptuelle de l'approche : elles montrent qu'un raisonnement fondé sur des structures schématiques peut être implémenté sous forme de programmes exécutables et appliqué à des tâches de déduction logique exprimées en langage naturel. Le résultat principal n'est donc pas seulement la performance obtenue, mais le fait qu'une même représentation puisse à la fois être inspectable, manipulable par un solveur, et reliée à des principes cognitifs de compréhension du langage.

Cette perspective suggère également qu'un langage schématique relativement simple peut servir de base commune pour représenter des situations variées. Les prédicats spatiaux utilisés possèdent en effet une sémantique géométrique fixée indépendamment du modèle de langage, ce qui réduit la charge imposée à celui-ci, puisqu'il n'a plus à apprendre ou à reconstruire le sens des prédicats dans chaque programme généré. Cette propriété contraste avec d'autres approches neurosymboliques reposant sur des prédicats définis *ad hoc* pour chaque problème [20, 4], dont la généralisation à de nouveaux domaines peut alors dépendre plus fortement de la capacité du modèle de langage à produire des prédicats correctement interprétables.

L'approche bénéficie également d'un ancrage formel solide, dans la mesure où la sémantique logique sous-jacente aux programmes ASP, à savoir la logique de *Here-and-There* (HT), a fait l'objet de nombreux travaux et a été étendue avec contraintes dans la logique HTc. L'étude de HTc avec des contraintes spatiales a été proposée dans [18], établissant ainsi le cadre formel sous-jacent aux programmes d'Embodied-LM. Plus récemment, la logique HTc a été enrichie par des opérateurs temporels [2], fournissant ainsi un cadre naturel pour caractériser des représentations schématiques dynamiques, où les objets, leurs paramètres et les relations qui les lient peuvent évoluer au cours du temps.

Les prochains axes de développement consistent, d'une part, à évaluer le système sur d'autres jeux de données où un raisonnement statique suffit, et, d'autre part, à exploiter cette possibilité d'extension dynamique afin de traiter des problèmes de raisonnement tels que la *tour de Hanoï* ou encore le *Blocks World*. Cette même extension viserait également à faire d'Embodied-LM un cadre de psychologie computationnelle pour la modélisation du raisonnement humain [14, 15] : en ajoutant des mécanismes comme un focus attentionnel, le système pourrait servir à simuler non seulement les processus de résolution des participants, mais aussi certaines préférences de construction des représentations ou certains biais observés dans le raisonnement.

Références

[1] Mehul Bhatt, Jae Hee Lee, and Carl Schultz. CLP(QS) : A declarative spatial reasoning framework. In Max J. Egenhofer, Nicholas A. Giudice, Reinhard Moratz, and Michael F. Worboys, editors, *Spatial Information Theory - 10th International Conference, COSIT 2011, Belfast, ME, USA, September 12-16, 2011. Proceedings*, volume 6899 of *Lecture Notes in Computer Science*, pages 210–230. Springer, 2011.

[2] Pedro Cabalar, Martín Diéguez, François Olivier, Torsten Schaub, and Igor Stephan. Towards constraint temporal answer set programming. *Theory and Practice of Logic Programming*, 2025.

[3] Francesco Calimeri, Wolfgang Faber, Martin Gebser, Giovambattista Ianni, Roland Kaminski, Thomas Krennwallner, Nicola Leone, Marco Maratea, Francesco Ricca, and Torsten Schaub. Asp-core-2 input language format. *Theory and Practice of Logic Programming*, 20(2) :294–309, 2020.

[4] Benjamin Callewaert, Simon Vandevelde, and Joost Vennekens. Verus-lm : a versatile framework for combining llms with symbolic reasoning, 2025.

[5] Leonardo De Moura and Nikolaj Bjørner. Z3 : An efficient smt solver. In *Proceedings of the 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, pages 337–340. Springer, 2008.

[6] Maria M Hedblom. *Image schemas and concept invention : cognitive, logical, and linguistic investigations*. Springer Nature, 2020.

[7] Mark Johnson. *The body in the mind : The bodily basis of reason and imagination*. Chicago : University of Chicago Press, 1987.

[8] Roland Kaminski, Javier Romero, Torsten Schaub, and Philipp Wanko. How to build your own asp-based system ?! *Theory and Practice of Logic Programming*, 23(1) :299–361, 2023.

[9] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

[10] George Lakoff and Rafael Núñez. *Where mathematics comes from*, volume 6. New York : Basic Books, 2000.

[11] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6) :517–540, 2024.

[12] Jean M Mandler and Cristóbal Pagán Cánovas. On defining image schemas. *Language and Cognition*, 6(4) :510–532, 2014.

[13] R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew D. Hardy, and Thomas L. Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences of the United States of America*, 121(41) :e2322420121, 2024.

[14] François Olivier. A logical framework for spatial mental models. In J. Šķilters, N. Newcombe, and D. Uttal, editors, *German Conference on Spatial Cognition XII*, pages 268–280. Springer, 2020.

[15] François Olivier. *Spatial relations in reasoning : a computational model*. PhD thesis, Université Paris sciences et lettres, 2022.

[16] François Olivier and Zied Bouraoui. Grounding agent reasoning in image schemas : A neurosymbolic approach to embodied cognition. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*, pages 2875–2879. IFAAMAS, 2025.

[17] François Olivier and Zied Bouraoui. Towards a neurosymbolic reasoning system grounded in schematic representations. In *Proceedings of the 19th International Conference on Neurosymbolic Learning and Reasoning (NeSy 2025)*, 2025.

[18] François Olivier and Carl Schultz. Here-and-there with constraints for spatial reasoning. In *Proceedings of the Sixth Workshop on Trends and Applications of Answer Set Programming (TAASP'23)*, 2023.

[19] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*, 2023.

[20] Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-lm : Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics : EMNLP 2023*, pages 3806–3824, 2023.

[21] Atsushi Shimojima. *Semantic Properties of Diagrams and Their Cognitive Potentials*. CSLI Publications, Stanford, California, 2015.

[22] A Srivastava, A Rastogi, A Rao, AAM Shoeb, A Abid, A Fisch, AR Brown, A Santoro, A Gupta, A Garriga-Alonso, et al. Beyond the imitation game : Quantifying and extrapolating the capabilities of language models (arxiv : 2206.04615). arxiv, 2023.

[23] Ronen Tamari, Chen Shani, Tom Hope, Miriam R L Petruck, Omri Abend, and Dafna Shahaf. Language (re)modelling : Towards embodied language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6268–6281, Online, July 2020. Association for Computational Linguistics.

[24] Przemysław Andrzej Wałęga, Carl Schultz, and Mehul Bhatt. Non-monotonic spatial reasoning with answer set programming modulo theories. *Theory and Practice of Logic Programming*, 17(2) :205–225, 2017.

SESSION 6 : Paradigme multi-agents présidée parElsy Kaddoum

Fusion de détecteurs d'objets par coordination multi-agent et a priori causaux

Alaa Daoud[1], alaa.daoud@uphf.fr
Hiba Alqasir[1], chloéhiba.alqasir@uphf.fr
Maxime Guèriau[2], maxime.gueriau@insa-rouen.fr
Corentin Bunel[3] corentin.bunel@cerema.fr

[1] Université Polytechnique Hauts-de-France, CNRS, UMR 8201-LAMIH, F-59313, Valenciennes, France
[2] INSA Rouen Normandie, LITIS UR 4108, F-76000 Rouen,France
[3] CEREMA, ENDSUM, Le Grand-Quevilly, France

Résumé

Les ensembles de détecteurs d'objets nécessitent une étape de fusion pour agréger des boîtes englobantes redondantes ou conflictuelles. Les approches classiques, telles que Weighted Boxes Fusion (WBF), reposent sur des seuils fixes d'Intersection over Union (IoU) et sur des scores de confiance homogènes, ce qui limite leur adaptabilité, leur décentralisation et leur interprétabilité. Nous proposons deux contributions complémentaires : (i) une reformulation de la fusion comme un problème de coordination multi-agent distribuée, où chaque boîte englobante devient un agent autonome capable de coopération, de compétition et de négociation ; (ii) une pondération par fiabilité comme connaissance causale a priori, appliquée après agrégation spatiale et associée à un seuil d'IoU adaptatif interprété comme intervention contextuelle, séparant consensus géométrique et confiance épistémique. La fusion est ainsi interprétée comme un mécanisme de coordination régulé par des a priori épistémiques et des interventions contextuelles, reliant systèmes multi-agents et raisonnement causal pour une perception adaptative et interprétable.

Mots-clés

Détection d'objets, fusion adaptative, boîtes englobantes, systèmes multi-agents, coordination, a priori causaux.

Abstract

Object detection ensembles require fusion to resolve redundant or conflicting bounding boxes. Classical methods such as Weighted Boxes Fusion (WBF) rely on fixed IoU thresholds and uniform detector trust, limiting adaptability, decentralization, and interpretability. Two complementary contributions are proposed : (i) a reformulation of fusion as a distributed multi-agent coordination problem (AWBF), where each bounding box becomes an autonomous agent capable of cooperation, competition, and negotiation ; (ii) a reliability-aware extension (AR-WBF) modeling detector reliability as a causal prior applied after spatial aggregation, with an adaptive IoU threshold acting as a contextual intervention that separates geometric consensus from epistemic trust. Ensemble fusion is thus interpreted as a layered coordination mechanism regulated by epistemic priors and contextual interventions, bridging multi-agent systems and causal reasoning for adaptive and interpretable perception.

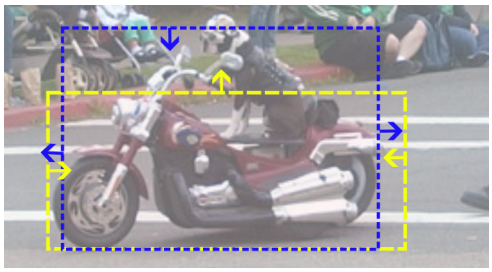


FIGURE 1 – Fusion adaptative de 2 boîtes englobantes sur une moto, image #495357 du jeu de données COCO

Keywords

Object detection, bounding-boxes, adaptive fusion, multi-agent systems, coordination, causal priors.

1 Problématique

La détection d'objets constitue un composant central des systèmes de perception embarqués, notamment en conduite autonome et en robotique. L'usage d'ensembles de détecteurs améliore la robustesse mais génère des prédictions multiples pour un même objet. Une étape de fusion devient alors nécessaire.

Les méthodes classiques telles que *Non-Maximum Suppression* (NMS) [5] ou Weighted Boxes Fusion (WBF) [6] réalisent une agrégation statique basée sur des seuils fixes d'IoU et des scores de confiance. NMS repose sur la suppression, tandis que WBF relève de la fusion. Sa supériorité sur NMS étant déjà établie dans [6], nous le prenons comme référence pour étudier des mécanismes de fusion. Cette approche soulève néanmoins plusieurs limites : 1. absence d'adaptation au contexte (taille, densité, variabilité des objets) ; 2. absence de modélisation explicite de la fiabilité relative des détecteurs ; 3. centralisation du processus décisionnel ; 4. faible interprétabilité des décisions de fusion. Ces constats nous conduisent aux questions suivantes :

RQ1 : Peut-on formuler la fusion de boîtes englobantes comme une coordination multi-agents distribuée permettant une décision adaptative et décentralisée ?

RQ2 : Comment intégrer explicitement la fiabilité des détecteurs en séparant consensus spatial et confiance épistémique, avec une interprétation causale de la fusion ?

2 Agentification de la fusion

Pour répondre à RQ1, nous proposons *Agentified WBF* (AWBF) [2], qui reformule la fusion comme un problème de coordination multi-agents adaptative (figure 1). Chaque boîte est incarnée par un agent autonome caractérisé par sa localisation, son score de confiance et sa capacité d'interaction locale.

2.1 Formalisation des interactions

Chaque agent propose une boîte englobante et peut interagir avec ses voisins en adoptant des comportements de compétition ou de coopération. Contrairement aux approches centralisées basées sur une métrique symétrique (IoU), nous utilisons une métrique asymétrique d'Intersection over Box (IoB), permettant de modéliser des relations directionnelles d'influence. L'IoB est définie par :

$$\text{IoB}(b_i, b_j) = \frac{|b_i \cap b_j|}{|b_i|},$$

ce qui permet de distinguer la force d'un agent i sur j de celle de j sur i.
Nous définissons alors une interaction compétitive entre deux agents A et B via leurs forces d'attaque et de défense :

$$S_{\text{attack}}(A, B) = s_A \cdot \text{IoB}(b_B, b_A),$$

$$S_{\text{defense}}(B, A) = s_B \cdot \text{IoB}(b_A, b_B),$$

où s_A et s_B désignent les scores de confiance des agents.
Le résultat de l'interaction est donné par :

$$R(A, B) = S_{\text{attack}}(A, B) - S_{\text{defense}}(B, A).$$

Le comportement collectif est régulé par un seuil T :

- si $R(A, B) > T$, A domine et B est supprimé ;
- si $R(A, B) < -T$, B domine et A est supprimé ;
- sinon, les deux agents coopèrent et fusionnent leurs boîtes.

Le seuil T, fixé empiriquement, contrôle le niveau de coopérativité : des valeurs élevées favorisent la fusion, tandis que des valeurs faibles induisent des interactions plus compétitives. Les interactions sont réalisées localement au sein de clusters définis par recouvrement spatial.

2.2 Cadre général de coordination.

Plus généralement, la compétition entre agents peut être formulée comme un mécanisme d'enchères. Chaque agent i soumet une offre fondée sur une fonction d'utilité U_i combinant confiance, recouvrement spatial et fiabilité, et propose une enchère pondérée par recouvrement $\text{Bid}_{i \to j}$:

$$\text{Bid}_{i \to j} = U_i \cdot \text{IoB}(b_j, b_i).$$

$$U_i = \lambda_1 s_i + \lambda_2 \text{IoU}(b_i, b_j) + \lambda_3 r_i,$$

où $\lambda_1, \lambda_2, \lambda_3$ sont des coefficients pondérant respectivement la confiance, le recouvrement spatial et la fiabilité du détecteur.

Le gagnant est déterminé de manière décentralisée par comparaison des enchères, tandis que des valeurs proches (contrôlées par T) conduisent à une fusion coopérative. Cette formulation unifie les mécanismes d'attaque-défense et permet d'intégrer explicitement des facteurs de fiabilité dans la décision.

Dans ce cadre, trois mécanismes de coordination ont été instanciés et évalués : (a) une reproduction décentralisée de WBF par moyenne pondérée locale ; (b) un mécanisme compétitif d'attaque–défense fondé sur l'IoB ; (c) une négociation multi-tour permettant un ajustement itératif jusqu'à un consensus local.

Cette reformulation transforme la fusion en un processus décisionnel décentralisé et adaptatif, où la dynamique de consensus émerge des interactions locales. La fiabilité des détecteurs demeure toutefois implicite, car elle n'est pas distinguée formellement de la confiance locale.

3 Fiabilité comme prior causal

Pour répondre à RQ2, nous proposons une extension fiabiliste de WBF, appelée *Adaptive Reliability WBF* (AR-WBF) [1], reposant sur deux mécanismes.

3.1 Pondération différée par fiabilité

Nous définissons un prior de fiabilité $r_k \in [0, 1]$ pour chaque détecteur M_k, dérivé de sa performance validée (mAP normalisée). La fusion produit une boîte agrégée b^* et un score de confiance associé s^*, la fiabilité intervenant uniquement dans la calibration finale du score :

$$b^* = \frac{\sum s_i b_i}{\sum s_i}$$

$$s^* = \left(\frac{1}{n} \sum s_i \right) \times (\alpha + \beta r_k)$$

où b_i et s_i désignent respectivement la boîte englobante et le score de confiance de la détection i, et n le nombre de détections dans le cluster.

Les paramètres α et β contrôlent respectivement le poids du consensus moyen et l'influence du prior de fiabilité ($\alpha + \beta = 1$). Cette paramétrisation rend explicite le compromis entre confiance collective et crédibilité épistémique.

En pratique, r_k est estimé à partir du mAP du détecteur M_k sur un jeu de validation, normalisé dans $[0, 1]$.

$$r_k = \frac{\text{mAP}_k - \min_j(\text{mAP}_j)}{\max_j(\text{mAP}_j) - \min_j(\text{mAP}_j) + \epsilon}.$$

Dans une lecture causale, r_k agit comme un prior modulant la crédibilité des détections issues de M_k, indépendamment du consensus spatial, ce qui permet de dissocier perception géométrique et confiance épistémique. Cette pondération différée évite qu'un détecteur peu fiable n'influence la géométrie du consensus et limite la suppression excessive, notamment dans des ensembles hétérogènes.

3.2 Intervention contextuelle via IoU

En notant σ_A la variance normalisée des tailles d'objets dans le cluster, nous définissons le seuil adaptatif :

$$\tau(c, A) = \tau_0 + \alpha_c \sigma_A.$$

Ce mécanisme agit comme une intervention sur la frontière décisionnelle : les petits objets ou contextes incertains déclenchent une fusion plus permissive, tandis que les objets stables imposent une contrainte plus stricte. La fusion devient ainsi sensible au contexte.

4 Experimentation

Nous évaluons AWBF sur le jeu de données COCO [3] en reprenant le benchmark complet proposé dans [6], comprenant 21 détecteurs et leurs prédictions associées. Ce protocole permet de comparer AWBF à WBF dans un cadre décentralisé, tout en conservant les mêmes entrées et conditions expérimentales. Les différentes variantes de AWBF — fusion coopérative (AWBF-f), compétition (AWBF-c) et négociation multi-tour (AWBF-n) — sont ensuite évaluées afin d'analyser l'impact des mécanismes de coordination sur le processus de fusion. Les performances sont mesurées à l'aide des métriques standards du protocole COCO, notamment la précision moyenne (AP) et le rappel moyen (AR), évalués à différents seuils IoU et pour plusieurs tailles d'objets (petits, moyens et grands).

TABLE 1 – Performances de WBF et des variantes AWBF sur le benchmark COCO, évaluées en AP et AR selon la taille des objets.

	Précision moyenne (AP)			Rappel moyen (AR)		
	Petits	Moyens	Grands	Petits	Moyens	Grands
WBF	0.605	0.731	**0.846**	**0.800**	**0.850**	**0.867**
AWBF-f	0.610	0.766	0.675	0.664	0.706	0.819
AWBF-c	0.322	0.636	0.632	0.653	0.743	0.840
AWBF-n	**0.622**	**0.780**	0.701	0.684	0.741	0.859

Les résultats montrent que les variantes AWBF peuvent améliorer la précision moyenne pour les petits et moyens objets, en particulier avec AWBF-n, qui obtient les meilleurs scores AP sur ces deux catégories. En revanche, WBF conserve un avantage net pour les grands objets ainsi que pour l'ensemble des métriques de rappel. Ces résultats suggèrent que la coordination adaptative peut favoriser la précision de certaines détections, mais qu'elle tend également à réduire la couverture globale des objets détectés par rapport à l'agrégation déterministe de WBF.

La figure 2 illustre l'évolution des boîtes englobantes au cours du processus de fusion adaptative, à trois étapes successives : état initial, intermédiaire et final, sur des images issues du jeu de données COCO. Ces exemples mettent en évidence les ajustements successifs produits par les mécanismes de coordination lors de la construction du consensus spatial.

Afin d'analyser finement l'impact de la fiabilité sur le processus de **AR-WBF**, nous adoptons un protocole contrôlé



FIGURE 2 – Exemples d'évolution des boîtes englobantes sous AWBF aux différentes étapes du processus de fusion adaptative (initiale, intermédiaire et finale) sur des images de COCO.

TABLE 2 – Performances de WBF et AR-WBF sur le jeu de validation COCO avec YOLOv5x et EfficientDet-B2, évaluées en AP et AR selon la taille des objets.

	Précision moyenne (AP)			Rappel moyen (AR)		
	Petits	Moyens	Grands	Petits	Moyens	Grands
WBF	0.281	0.453	0.547	0.481	0.676	0.785
AR-WBF	0.281	**0.458**	**0.553**	**0.503**	**0.701**	**0.801**

en sélectionnant deux détecteurs issus du benchmark, plutôt que l'ensemble complet des modèles. Les expériences sont menées sur un sous-ensemble de 500 images du jeu de validation COCO, sélectionnées aléatoirement.

Nous retenons **YOLOv5x** [4] et **EfficientDet-B2** [7], représentant des architectures complémentaires : YOLOv5x favorisant le rappel, et EfficientDet-B2 privilégiant la précision. Ce choix permet d'isoler les effets de la pondération par fiabilité dans un cadre hétérogène mais interprétable.

Les prédictions de ces modèles sont évaluées individuellement puis fusionnées via WBF et AR-WBF. Le prior de fiabilité r_k est estimé à partir du mAP sur un sous-ensemble de validation COCO, puis normalisés avant la fusion. Pour assurer une comparaison équitable, tous les modèles sont pondérés uniformément et un seuil IoU de base de 0.5 est utilisé. Les performances sont évaluées selon le protocole COCO (AP et AR).

La table 2 présente les résultats obtenus avec WBF et AR-WBF sur les prédictions de YOLOv5x et EfficientDet-B2. AR-WBF obtient des performances comparables à WBF pour les petits objets, tout en améliorant systématiquement les métriques AP et AR pour les objets moyens et grands. Les gains observés en rappel suggèrent que l'introduction d'une pondération par fiabilité permet de mieux préserver les détections pertinentes lors du processus de fusion, tout en maintenant une cohérence spatiale entre les boîtes agrégées.

Figure 3 compare les résultats de fusion entre WBF et AR-WBF sur des échantillons sélectionnés à partir des prédic-

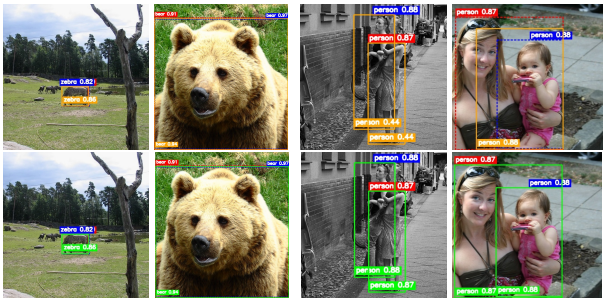


FIGURE 3 – Similarité (à gauche) et divergence (à droite) entre WBF (en haut) et AR-WBF (en bas) sur des exemples issus de YOLOv5x et EfficientDet-B2.

tions de YOLOv5x (bleu) et EfficientDet-B2 (rouge). Dans de nombreux cas, les deux méthodes produisent des résultats similaires. Cependant, des divergences apparaissent lorsque l'IoU se situe à proximité du seuil de fusion, conduisant à des décisions d'agrégation différentes entre WBF et AR-WBF.

Lorsque le recouvrement entre boîtes se situe à proximité du seuil de fusion, WBF peut produire des décisions d'agrégation instables, soit en fusionnant des détections spatialement discordantes, soit en conservant des duplications. AR-WBF réduit ces effets grâce à un seuil IoU adaptatif et à une pondération par fiabilité appliquée après regroupement, ce qui permet d'obtenir une fusion plus cohérente spatialement.

5 Synthèse et perspectives

Les résultats expérimentaux présentés dans la section précédente mettent en évidence deux dynamiques complémentaires dans le processus de fusion.

D'une part, AWBF introduit une coordination distribuée entre détecteurs afin de reproduire le comportement de WBF dans un cadre décentralisé. Les résultats montrent que certaines variantes, en particulier AWBF-n, améliorent la précision moyenne pour les petits et moyens objets (cf. table 1). Ces gains suggèrent que les mécanismes de coordination peuvent faciliter l'agrégation de détections partiellement redondantes. En revanche, WBF conserve de meilleures performances sur les objets de grande taille ainsi que sur plusieurs métriques de rappel, indiquant que les interactions entre agents peuvent également introduire une variabilité dans le consensus spatial produit lors de la fusion.

D'autre part, AR-WBF introduit une pondération des détecteurs fondée sur leur fiabilité estimée. Les résultats présentés dans la table 2 montrent des améliorations cohérentes des métriques AP et AR pour les objets moyens et grands. Cette tendance est également visible dans l'analyse des gains relatifs présentée dans la table 3. Contrairement à AWBF, dont les performances varient fortement selon les métriques considérées, AR-WBF présente des gains plus homogènes sur les indicateurs liés à la précision de localisation et au rappel.

TABLE 3 – Gains relatifs (%) obtenus par AWBF et AR-WBF par rapport à WBF sur différentes métriques AP et AR du protocole COCO.

	AP_{75} (%)	AP_M (%)	AR_{50} (%)	AR_{75} (%)
AWBF	−8.4	3.5	4.9	−10.1
AR-WBF	1.2	0.5	1.0	2.3

Ces deux approches répondent ainsi à des objectifs complémentaires. AWBF cherche à modéliser des mécanismes de coordination distribuée entre détecteurs, tandis que AR-WBF introduit une calibration globale fondée sur la fiabilité des modèles. Le processus de fusion peut ainsi être vu comme une combinaison entre interactions locales et pondération globale des prédictions.

Les mécanismes présentés s'appuient sur des résultats expérimentaux détaillés dans [2, 1], dont nous synthétisons ici les éléments essentiels afin de mettre en évidence la complémentarité entre coordination multi-agents et calibration par fiabilité.

AWBF introduit une coordination distribuée entre détecteurs, capable d'améliorer certaines métriques de précision au prix d'une plus grande variabilité du consensus spatial. À l'inverse, AR-WBF stabilise le processus de fusion grâce à une pondération globale fondée sur la fiabilité des modèles. Ces observations suggèrent qu'une combinaison entre coordination locale et calibration par fiabilité constitue une direction prometteuse pour la fusion multi-sources.

Références

[1] Alaa Daoud and Hiba Alqasir. From consensus to causality : Adaptive reliability fusion for object detection ensembles. In *CALM-26*, 04 2026.

[2] Alaa Daoud, Corentin Bunel, and Maxime Guériau. Introducing multiagent systems to av visual perception sub-tasks : A proof-of-concept implementation for bounding-box improvement. In *ATT-24 @ ECAI*, 2024.

[3] Tsung-Yi Lin et al. Microsoft coco : Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[4] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, et al. Yolov5 by ultralytics, 2023. YOLOv5x model variant.

[5] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *ICPR'06*, volume 3, pages 850–855. IEEE, 2006.

[6] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion : Ensembling boxes from different object detectors. *Image and Vision Computing*, 2021.

[7] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet : Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

Approche par optimisation sous contraintes distribuée Divide-and-Coordinate pour le transport urbain multimodal

Laura Trujillo[1,2], Elsy Kaddoum[1], Gauthier Picard[2]

[1] Univ. Toulouse, Institut de recherche en informatique de Toulouse (IRIT), France
[2] ONERA/DTIS, Université de Toulouse, France

16 février 2026

Résumé

La Mobilité Servicielle (MaaS) vise à répondre aux enjeux de congestion, de dépendance à l'automobile et de transition écologique en intégrant plusieurs services de transport au sein d'une plateforme unifiée de planification d'itinéraires personnalisés. La planification réaliste d'itinéraires multimodaux demeure toutefois un défi majeur : les approches existantes considèrent peu de modes, négligent les ressources de mobilité personnelles et reposent principalement sur des schémas centralisés. Nous introduisons le Personalized and Resource-aware Multimodal Routing Problem *(PRMRP), une formalisation de la planification multimodale et multi-utilisateurs. Deux formulations multi-agents sont proposées, basées sur l'optimisation distribuée sous contraintes (DCOP) et une approche* Divide-and-Coordinate *(DaC), et évaluées sur des scénarios réalistes à l'échelle d'une ville.*

Mots-clés

Mobilité servicielle, Transports, Optimisation Distribuée, Systèmes Multi-Agents

Abstract

Urban mobility systems face growing challenges related to congestion, car dependency, and ecological transition. Mobility as a Service (MaaS) addresses these issues by integrating multiple transportation services within a unified personalized trip planning platform. However, realistic multimodal itinerary planning remains one of the central challenges. Existing approaches typically consider a limited set of transport modes, rarely account for personal mobility resources, and rely on centralized schemes that decouple service composition from resource allocation. We introduce the Personalized and Resource-aware Multimodal Routing Problem *(PRMRP), a formal formulation for multi-modal and multi-user itinerary planning with resource allocation. We propose two multi-agent formulations for the itinerary planning component : a Distributed Constraint Optimization (DCOP) model that could be solved by dedicated algorithms and a Divide-and-Coordinate (DaC) model based on the dual decomposition of the problem into subproblems to coordinate. We evaluate both approaches against a centralized complete solver based on a MILP formulation and a greedy heuristic baseline, on realistic city-scale multi-modal transportation scenarios.*

Keywords

Mobility-as-a-Service, Transportation, Distributed Optimization, Multi-Agent Systems

1 Introduction

Les zones urbaines font face à des défis croissants liés à la congestion, à la dépendance à l'automobile et à la transition écologique. Dans ce contexte, la mobilité servicielle (Mobility as a Service, MaaS) s'impose comme un paradigme prometteur en intégrant plusieurs services de mobilité au sein d'une plateforme unifiée de planification et de réservation de trajets personnalisés [16]. La mise en œuvre effective du MaaS repose notamment sur la recommandation d'itinéraires multimodaux et multi-utilisateurs, qui reflète pleinement ce paradigme en intégrant des services émergents tels que le transport à la demande ou le covoiturage [25, 30]. Cependant, les approches existantes restent limitées : elles considèrent un petit sous ensemble de modes, souvent issus d'un unique opérateur, offrent une personnalisation restreinte (souvent temporelle uniquement) et relèguent les ressources de mobilité personnelles au rôle d'options de secours. Elles s'appuient majoritairement sur des formulations centralisées en programmation linéaire en nombres entiers mixtes (MILP), dont la complexité croît rapidement avec le nombre d'utilisateurs, de modes et d'arcs temporels, dépassant les capacités des solveurs. Ces approches requièrent en outre une vision globale des données privées des utilisateurs, soulevant des enjeux de confidentialité et de conformité au RGPD [9], et induisent une latence incompatible avec des scénarios MaaS réalistes. Enfin, la séparation récurrente entre allocation des ressources et composition modale peut conduire à des solutions sous-optimales. Sur la base de cette analyse, nous défendons une approche décentralisée de la planification multimodale intégrée dans le cadre du MaaS, fondée sur les contributions suivantes :

(i) l'introduction du *Personalized and Resource-aware Multimodal Routing Problem* (PRMRP), un cadre unifié permettant une composition flexible de tout mode de transport conditionnée par la disponibilité

des ressources et la prise en compte explicite des préférences utilisateurs ;
(ii) une formulation distribuée de la composante de routage du PRMRP sous forme de problème d'optimisation sous contraintes (DCOP) ;
(iii) une adaptation de cette formulation dans une approche *Divide-and-Coordinate* (DaC), résolue par un algorithme dédié, DaCSA-PRMRP ;
(iv) une évaluation expérimentale comparant DaCSA-PRMRP à des algorithmes DCOP, à un programme linéaire en nombres entiers centralisé et à une heuristique gloutonne.

La section 2 présente les approches classiques et distribuées de la planification d'itinéraires multimodaux. La section 3 introduit le modèle PRMRP. La section 4 décrit la formulation DCOP et l'approche DaC proposées. La section 5 détaille le protocole expérimental et les résultats obtenus, avant que la section 6 ne conclue et esquisse des perspectives de recherche.

2 Travaux antérieurs

Transport multimodal Les problèmes de planification des transports peuvent être caractérisés selon deux dimensions : (i) le nombre d'utilisateurs impliqués (mono- ou multi-utilisateur) et (ii) le nombre de modes de transport intégrés (mono- ou multimodal), dont le croisement définit quatre grandes familles de problèmes. Les travaux sur la planification d'itinéraires ont progressivement évolué de formulations mono-modales vers des approches multimodales plus intégrées. Les premiers modèles traitaient le cas mono-utilisateur et mono-modal à travers des problèmes classiques de plus court chemin sur des réseaux à une seule couche [7, 4]. Des extensions ultérieures ont considéré le cas mono-utilisateur multimodal en combinant des réseaux hétérogènes (route, rail, vélo, marche) [27, 2], permettant une planification porte-à-porte réaliste via des représentations dépendantes du temps ou de la fréquence [3, 6], mais sans coordination entre requêtes.

Des approches plus récentes ont introduit des perspectives multi-utilisateurs, notamment via des problèmes de gestion de flottes mono-modales tels que le Transport à la Demande [5, 23, 12] et les services de transport partagé [20], qui optimisent le routage, l'affectation des passagers et la planification sous contraintes de capacité et de fenêtres temporelles. Ces approches coordonnent les utilisateurs au sein d'un mode unique. Des formulations pleinement multimodales et multi-utilisateurs ont ensuite été proposées en combinant des modèles de routage centrés utilisateur avec des services de mobilité partagée [15, 25, 18], en étendant les modèles classiques de Transport à la Demande par l'intégration de transferts intermodaux et des contraintes de partage de ressources.

Malgré ces avancées, plusieurs limites persistent. L'intégration modale demeure souvent partielle, restreinte à un nombre réduit de modalités [15], tandis que les véhicules personnels sont généralement traités comme des solutions de secours plutôt que comme des modes pleinement combinables [17]. Les séquences de modes admissibles sont fréquemment prédéfinies, souvent selon un paradigme premier/dernier kilomètre, ce qui limite la flexibilité de la composition modale [17]. La personnalisation reste principalement temporelle ou budgétaire, avec une prise en compte limitée des préférences qualitatives [1, 31]. Enfin, les interactions entre utilisateurs sont le plus souvent confinées à des services isolés et la coordination de multiples itinéraires dans le temps et à travers plusieurs services successifs demeure peu explorée.

Méthodes centralisées La majorité des modèles multimodaux et multi-utilisateurs reposent sur des formulations centralisées, généralement sous forme de programmes linéaires en nombres entiers ou mixtes (ILP/MILP) [29, 25]. Le passage à l'échelle est abordé via des phases de génération de candidats ou des stratégies de décomposition [17], qui dissocient souvent la composition modale et l'allocation des ressources, au risque de produire des solutions globalement sous-optimales. Pour améliorer la scalabilité, des heuristiques et métaheuristiques issues du Transport à la Demande sont couramment utilisées, telles que des stratégies gloutonnes, des heuristiques d'insertion, des algorithmes génétiques, l'optimisation par essaims particulaires [1] ou la recherche de voisinage [21], en échange d'un compromis entre optimalité et efficacité computationnelle. Ces approches supposent toutefois une information complète et un décideur central unique, ce qui limite leur applicabilité dans des contextes MaaS dynamiques et décentralisés et soulève des enjeux de confidentialité [14].

Méthodes décentralisées Les problèmes de recommandation d'itinéraires impliquent des entités autonomes (utilisateurs, fournisseurs de services) disposant de connaissances locales et gérant leurs propres ressources. Ils relèvent naturellement des systèmes multi-agents. Des cadres décentralisés permettent une prise de décision locale, la préservation de la confidentialité [26] et une meilleure adaptation aux environnements dynamiques, mais restent encore peu explorés pour la planification multimodale. Dans [8], les auteurs montrent qu'une approche multi-agents peut faire passer à l'échelle la planification de trajets co-modaux par la coordination de décisions locales, sans toutefois s'appuyer sur une optimisation distribuée formelle. Des cadres décentralisés modélisent explicitement la coopération entre agents, c'est le cas du formalisme des problèmes d'optimisation distribuée sous contraintes (DCOP), qui permet de coordonner des décisions locales vers un objectif global [11]. Bien que les DCOP aient été appliqués à des problèmes de logistique et de routage, leur utilisation pour la planification de la mobilité multimodale n'existe pas à notre connaissance. Malgré un coût potentiel en communication, ils offrent des garanties d'optimalité et une représentation naturelle de la coordination décentralisée, ce qui en fait une base prometteuse pour ce travail.

3 Définition du problème

Nous introduisons le *Personalized and Resource-aware Multimodal Routing Problem* (PRMRP), défini comme le

tuple $\langle \mathcal{Q}, \mathcal{N}, \mathcal{O}, \mathcal{S}, \mathcal{R} \rangle$, où $\mathcal{Q}$, $\mathcal{N}$, $\mathcal{O}$, $\mathcal{S}$ et $\mathcal{R}$ désignent respectivement les ensembles des requêtes, des nœuds de transfers, des opérateurs, des services et des ressources.

Requêtes utilisateurs Soit $\mathcal{U}$ l'ensemble des utilisateurs soumettant des requêtes. Chaque requête $q_i \in \mathcal{Q}$ est définie par :

$$q_i = (u_i, d_i, a_i, [t_i^{ED}, t_i^{LA}]),$$

où :

- $u_i \in \mathcal{U}$ est l'utilisateur associé, caractérisé par : (1) $pr_i = \left(pr_i^{\text{vehicles}}, pr_i^{\text{advantages}}\right)$, représentant ses ressources de mobilité personnelles (p. ex. véhicules possédés) et ses avantages liés à des abonnements (p. ex. train, bus) ; (2) $w_i \in \mathbb{R}^k_{\geq 0}$, un vecteur de préférences normalisé pondérant les critères d'optimisation (p. ex. temps, coût, confort, émissions) ; et (3) β_i, l'ensemble des contraintes spécifiques à l'utilisateur (p. ex. accessibilité, distance de marche, restrictions modales) ;
- $d_i, a_i \in \mathcal{N}$ sont respectivement les points de départ et d'arrivée ;
- $[t_i^{\text{ED}}, t_i^{\text{LA}}]$ correspond aux instants de départ au plus tôt (ED) et d'arrivée au plus tard (LA).

Services Soit $\mathcal{S} = \{s_1, \ldots, s_k, \ldots, s_{|\mathcal{S}|}\}$ l'ensemble des services de transport disponibles. Chaque service $s_k \in \mathcal{S}$ est géré par un opérateur unique. Pour chaque requête q_i, $S_i \subseteq \mathcal{S}$ désigne l'ensemble des services accessibles.

Opérateurs Soit $\mathcal{O} = \{o_1, \ldots, o_p, \ldots, o_{|\mathcal{O}|}\}$ l'ensemble des opérateurs de mobilité. Chaque opérateur $o_p \in \mathcal{O}$ gère un ou plusieurs services de transport et détermine ceux accessibles à chaque requête.

Nœuds Un noeud repésente un point d'entrée ou de sortie pour un ou plusieurs services. Chaque nœud $n \in \mathcal{N}$ peut avoir des caractéristiques opérationnelles selon le type de service qui s'y trouve. On note $\kappa^{\text{out}}_{n,s}$ et $\kappa^{\text{in}}_{n,s}$ respectivement le nombre maximal de départs et d'arrivées simultanés autorisés au nœud n pour le service $s \in \mathcal{S}$. Ces paramètres modélisent des goulots d'étranglement opérationnels tels que des capacités limitées de quais, de zones de chargement ou de rotation des véhicules. Certains services pouvant être restreints à des requêtes spécifiques (p. ex. véhicules privés) ou exclus par un utilisateur, chaque requête $q_i \in \mathcal{Q}$ dispose de son propre ensemble de nœuds de correspondance accessibles N_i. L'ensemble global des nœuds de correspondance est alors défini par :

$$\mathcal{N} = \bigcup_{i=1}^{|\mathcal{Q}|} N_i.$$

On compte aussi parmi N_i, les points de départ et d'arrivé de q_i

Ressources Pour chaque service $s_k \in \mathcal{S}$, $\mathcal{R}_k = \{r_1, \ldots, r_\ell, \ldots, r_{|\mathcal{R}_k|}\}$ est l'ensemble des ressources associées. Chaque ressource $r_\ell \in \mathcal{R}_k$ représente une unité élémentaire de transport, de capacité $cap(r_\ell) \in \mathbb{N} \cup \{+\infty\}$. Les ressources disposant d'horaires prédéfinis sont dites *fixes*, tandis que les autres sont dites *libres*.

A partir de ces définitions, nous pouvons définir la notion d'itinéraire et l'objectif du problème, en passant par les concepts de tronçon candidat et de tronçon instancié.

Tronçon candidat Pour chaque requête $q_i \in \mathcal{Q}$, soit $L_i = \{l_1, \ldots, l_m, \ldots, l_{|L_i|}\}$ l'ensemble des tronçons candidats accessibles. Un tronçon candidat correspond à un trajet potentiel, opéré par un service entre deux noeuds. Formellement, chaque tronçon $l_m \in L_i$ est défini par $l_m = (d_m, a_m, s_m)$, où $d_m, a_m \in N^i$ sont les nœuds d'entrée et de sortie du service $s_m \in \mathcal{S}$.

Tronçon instancié L'affectation d'une ressource $r_\ell \in \mathcal{R}^k$ à un tronçon candidat $l_m \in L_i$ produit un tronçon instancié via l'application :

$$\psi : (L_i \times \mathcal{R}^k) \to \mathcal{D}_i,$$
$$\psi(l, r) = (t, dur, z)$$

où $\mathcal{D}_i$ ne contient que les tronçons instanciés pour les paires (l, r) faisables c.à.d. pour lesquelles la ressource effectivement peut opérer le service associé. Les variables t, dur et z désignent respectivement l'heure de départ, la durée et un vecteur de caractéristiques opérationnelles liées aux critères d'optimisation (p. ex. coût, énergie, confort). Ces données sont déduite de l'allocation de la ressource au tronçon.

Itinéraire Un itinéraire I_i associé à la requête q_i est alors sune séquence ordonnée de tronçons instanciés :

$$I_i = (\psi_0, \ldots, \psi_j, \psi_{j+1}, \ldots, \psi_{|I_i|-1})$$

Pour être faisable, il doit satisfaire la cohérence temporelle :

$$t_{j+1} \geq t_j + dur_j + \tau(a_j, s_j, s_{j+1}, u_i)$$

où τ est une fonction de temps de correspondance dépendant du nœud, des services et de l'utilisateur. D'autres contraintes de faisabilité pourraient s'ajouter à cette contrainte minimale, selon les modes choisis et la modélisation qui en est faite (contraintes de flux du point de départ à l'arrivée ou contraintes de temps minimum de marche entre deux tronçons successifs). On peut dire de façon générale que le coût d'un itinéraire est défini comme la somme des coûts des tronçons instanciés et des coûts de correspondance, y compris l'accès à l'origine et à la destination :

$$Cost_i^{\text{it}}(I_i) = Cost_i^{\text{corr}}(d_i, \psi_0) + \sum_{j=0}^{|I_i|-1} Cost_i^{\text{tronçon}}(\psi_j) + \sum_{j=0}^{|I_i|-1} Cost_i^{\text{corr}}(\psi_j, \psi_{j+1}) + Cost_i^{\text{corr}}(\psi_{|I_i|-1}, a_i) \quad (1)$$

où $Cost_i^{\text{tronçon}}(\cdot)$ est le coût scalaire, dépendant de l'utilisateur, associé à un tronçon instancié, défini par :

$$Cost_i^{\text{tronçon}}\big(\psi(l, r)\big) \;=\; \langle w_i, \mathbf{z} \rangle.$$

Le terme $Cost_i^{\text{corr}}(\cdot)$ modélise les coûts d'accès ou correspondance dont on connait le vecteur de caractéristiques z', agrégé en un coût par produit scalaire avec le même vecteur de préférences w_i.

$$Cost_i^{\text{corr}}\big(\psi_j, \psi_{j+1}\big) \;=\; \langle w_i, \mathbf{z}' \rangle.$$

Objectif principal L'objectif est de sélectionner des tronçons candidats via l'allocation de ressources spécifiques, de manière à construire des itinéraires faisables et optimisant avec précision les préférences des utilisateurs.

Solution faisable Une solution $\Phi : \mathcal{Q} \to \mathcal{I}$ associe à chaque requête q_i un itinéraire $I_i = \Phi(q_i)$. Une solution est faisable si tous les itinéraires indépendants sont faisables, si toutes les contraintes utilisateur β_i sont satisfaites et si aucun conflit de ressources n'apparaît à aucun instant.

4 Modèle DaC pour le PRMRP

Dans le PRMRP, pour une requête donnée, les décisions principales sont locales : elles concernent pour chaque créneau de l'itinéraire la sélection d'un tronçon et de son instant d'entrée. Le couplage intervient principalement aux points de correspondance, où des tronçons successifs doivent être temporellement cohérents. De plus, des dépendances sont induites entre itinéraires par le partage de ressources. Nous exploitons cette structure partiellement locale, en formalisant le PRMRP comme un DCOP et en l'adaptant pour une approche de résolution *Divide-and-Coordinate*. Dans ces approches, les décisions locales sont optimisées indépendamment, tandis que la coordination est imposée uniquement pour les décisions partagées. Cela permet une recommandation d'itinéraires scalable, centrée utilisateur et respectueuse de leur confidentialité. À ce stade, le modèle se concentre sur la logique centrale de composition des services et suppose des capacités de ressources infinies. Les contraintes explicites de partage des ressources et la coordination multi-utilisateurs sont laissées à des travaux futurs. Dans ce contexte, les requêtes sont indépendantes.

4.1 Modèle DCOP pour le PRMRP

Un problème d'optimisation distribuée sous contraintes (DCOP) [11] étend le problème d'optimisation sous contraintes classique (COP) à un cadre multi-agents, dans lequel des agents, disposant uniquement de connaissances locales, doivent se coordonner par communication afin de minimiser un coût global. Formellement, un DCOP est défini comme un tuple $P = \langle \mathcal{A}, \mathcal{X}, \mathcal{D}, \mathcal{C}, \mu, f \rangle$, où $\mathcal{A} = \{a_1, \ldots, a_m\}$ est l'ensemble des agents, $\mathcal{X} = \{x_1, \ldots, x_n\}$ l'ensemble des variables de décision à domaines finis $\mathcal{D} = \{D_{x_1}, \ldots, D_{x_n}\}$, $\mathcal{R} = \{r_1, \ldots, r_k\}$ un ensemble de relations associant des coûts dans $\mathbb{R} \cup \{\infty\}$ aux affectations de variables, $\mu : \mathcal{X} \to \mathcal{A}$ l'application associant chaque variable à un agent, et $f : \prod_i D_{x_i} \to \mathbb{R}$ la fonction objectif globale, généralement définie comme la somme des coûts de toutes les relations. Résoudre un DCOP consiste à trouver une affectation des variables minimisant f. De nombreux algorithmes existent pour résoudre des DCOP, allant de méthodes exactes [13, 22, 24] à des approches approximées [19, 10], qui échangent des garanties d'optimalité contre une meilleure scalabilité.

La formulation DCOP proposée pour le PRMRP est illustrée en Figure 1. Chaque requête utilisateur q_i est prise en charge par un agent utilisateur (rectangle transparent) chargé de construire un itinéraire à l'aide d'un nombre fixe

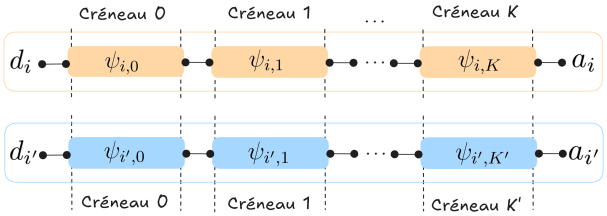


FIGURE 1 – Formulation DCOP avec décomposition des itinéraires en créneaux.

K (différent pour chaque q_i) d'agents créneau ordonnés (rectangles pleins). une variable $\psi_{i,k}$ est associée à chaque agent créneau. Elles permettent de sélectionner un tronçon instancié ou l'état inactif. La séquence de créneaux définit un itinéraire d'au plus K tronçons. Les dépendances entre décisions sont modélisées par des relations binaires partagées entre créneaux successifs (flèches circulaires). La figure illustre cette structure pour deux requêtes utilisateur, sans interaction entre elles.

L'ensemble complet des tronçons déjà instanciés est fourni au modèle de résolution puisque l'allocation de ressource n'est pas une décision. Nous supposons que les opérateurs sont capables de gérer et de prédire la disponibilité des ressources et leurs horaires. Pour une requête donnée, l'ensemble des tronçons instanciés est construit à partir des points de départ et d'arrivée, de la fenêtre temporelle et des services localement accessibles fournis par les opérateurs. La marche n'est pas considérée comme un service, car elle est réservée aux correspondances.

4.1.1 Domaines ($\mathcal{D}$)

Ensemble des tronçons instanciés disponibles pour la requête i, augmenté d'une valeur nulle $\perp$:

$$D_i^{\text{tronçon}} = \mathcal{D}_i \cup \{\perp\}.$$

4.1.2 Variables ($\mathcal{X}$)

Pour chaque requête i et chaque créneau $k \in \{1, \ldots, K\}$, une variable unique $\psi_{i,k} \in \mathcal{D}_i^{\text{tronçon}}$ est définie. Pour chaque variable, on note :

- $\ell_{i,k}$: le tronçon candidat associé au tronçon instancié ($\ell_{i,k} = \perp$ pour un tronçon nul) ;
- $t_{i,k}^{\text{in}}$: l'instant d'entrée du tronçon instancié (0 indique un créneau inactif). L'instant de sortie $t_{i,k}^{\text{out}}$ est déduit de $t_{i,k}^{\text{in}}$ et de la durée du tronçon.

4.1.3 Agents ($\mathcal{A}$)

Nous définissons un agent par créneau pour chaque requête :

$$\mathcal{A} = \{A_{i,k} \mid i \in \mathcal{Q},\ k = 1, \ldots, K\}.$$

Chaque agent $A_{i,k}$ contrôle $\psi_{i,k}$. Ainsi, chaque agent représente un créneau disponible pour un tronçon dans l'itinéraire d'une requête.

4.1.4 Contraintes et relations de coût ($\mathcal{C}$)

Les contraintes dures définissent les relations de faisabilité entre les variables de décision et restreignent l'ensemble des affectations admissibles. Les relations de coût ne sont évalués que sur les affectations faisables.

Contraintes Les contraintes dures sont exprimées sous forme d'ensembles caractérisant les décisions valides ; elles peuvent être traduites directement en relations DCOP équivalentes en attribuant un coût infini aux configurations invalides.

Activation du premier créneau : impose l'activation du premier créneau afin d'éviter des scénarios ne comportant que de la marche.

$$\mathcal{C}_{i,0}^{\text{premier_tronçon}} = \{\psi_{i,0} | \ell_{i,0} = \bot\}.$$

Compacité de l'itinéraire : un itinéraire doit être compact ; dès qu'un créneau est inactif, tous les créneaux suivants doivent l'être également ; retrouver le dernier créneau actif avec cette contrainte devient trivial.

$$\mathcal{C}_{i,k}^{\text{compact}} = \{(\psi_{i,k}, \psi_{i,k+1}) \mid \ell_{i,k} \neq \bot \ \vee \ \ell_{i,k+1} = \bot\}.$$

Temps minimal de correspondance : Soit la fonction $ok(a, b, t_1, t_2)$ retournant vrai si la correspondance est faisable entre deux noeuds i.e. :

$$t_2 \geq t_1 + \textit{min_transfer_time}(a, b)$$

On définit alors la contrainte de temps minimal de correspondance comme :

$$\mathcal{C}_{i,k}^{\text{min_corr}} = \left\{(\psi_{i,k}, \psi_{i,k+1}) \;\middle|\; \textit{condition}\right\}$$

avec $\textit{condition} = \big((\ell_{i,k} = \bot) \ \vee \ (\ell_{i,k+1} = \bot) \ \vee \ ok(a_{i,k}, d_{i,k+1}, t_{i,k}^{\text{out}}, t_{i,k+1}^{\text{in}})\big)$. Des contraintes supplémentaires de correspondance portent sur le choix des modes : deux tronçons consécutifs de même mode ne sont pas autorisés, sauf pour les transports publics avec des lignes distinctes.

Temps minimal de correspondance à l'origine : La correspondance entre le point de départ et le premier créneau doit être faisable :

$$\mathcal{C}_{i}^{\text{min_dep_corr}} = \{(\psi_{i,0} \mid ok(d_i, d_{i,0}, t_i^{\text{ED}}, t_{i,0}^{\text{in}})\}.$$

Temps minimal de correspondance vers la destination : La correspondance entre le créneau terminal et la destination doit être faisable. Un créneau k est dit *"terminal k"* si $\ell_{i,k} \neq \bot$ et $\ell_{i,k+1} = \bot$. Il est *"terminal K"* si $\ell_{i,k} \neq \bot$ et $k = K$. On définit alors :

$$\mathcal{C}_{i,k}^{\text{min_arr_corr}} = \{(\psi_{i,k}, \psi_{i,k+1}) \mid cond_k\}\,,$$

où $cond_k = \text{term}_\text{k}(\psi_{i,k}, \psi_{i,k+1}) \Rightarrow ok(a_{i,k}, a_i, t_{i,k}^{\text{out}}, t_i^{\text{LA}})$. Et pour le cas *terminal K* :

$$\mathcal{C}_{i,k}^{\text{min_arr_corr_K}} = \{\psi_{i,k} \mid cond_K\}\,.$$

où $cond_K = \text{term}_\text{K}(\psi_{i,k}) \Rightarrow ok(a_{i,k}, a_i, t_{i,k}^{\text{out}}, t_i^{\text{LA}})$.

Dans la suite du papier, nous feront référence aux relations qui représentent ces contraintes dures avec les notations : $R^{\text{premier_tronçon}}$, $R^{compact}$, R^{min_corr}, $R^{min_dep_corr}$ et $R^{min_arr_corr}$.

Relations de coût Dans la formulation DCOP, les relations de coût implémentent les composantes du coût d'itinéraire définies à la Section 3. Les coûts ne sont appliqués que lorsque les créneaux et correspondances sont actives et ils correspondent exactement au coût utilisateur $Cost_i^{\text{it}}$.

Coût de tronçon : Pour chaque requête i et créneau k, la sélection d'un tronçon non nul $\ell_{i,k} \neq \bot$ à l'instant $t_{i,k}^{\text{in}}$ engendre le coût du tronçon instancié correspondant :

$$R^{\text{tronçon}}(\psi_{i,k}) = \mathbb{I}[\ell_{i,k} \neq \bot]\ Cost_i^{\text{tronçon}}\big(\psi_{i,k}\big),$$

Coût de correspondance : Une correspondance entre deux créneaux consécutifs actifs engendre le coût suivant :

$$R^{\text{corr}}(\psi_{i,k}, \psi_{i,k+1}) = \begin{cases} 0 & \text{si } \textit{fin}, \\ Cost_i^{\text{corr}}(\psi_{i,k}, \psi_{i,k+1}) & \text{sinon.} \end{cases}$$

avec $\textit{fin} = \big((\ell_{i,k} = \bot) \ \vee \ (\ell_{i,k+1} = \bot)\big)$

Coût d'accès à l'origine : Le coût de correspondance entre le point de départ et le premier créneau actif est défini par :

$$R^{\text{dep}}(\psi_{i,0}) = Cost_i^{\text{corr}}(d_i, \psi_{i,0})$$

Coût d'accès à la destination : Le coût d'arrivée à la destination est modélisé par des relations binaires $R_i^{\text{arr}_k}$ et $R_i^{\text{arr}_K}$, qui n'interviennent que lorsque la fin d'itinéraire est détectée.

$$R_i^{\text{arr}}(\psi_{i,k}, \psi_{i,k+1}) = \mathbb{1}_{\text{term}_k(\psi_{i,k}, \psi_{i,k+1})}\ Cost_i^{\text{corr}}(\psi_{i,k}, a_i)$$

et, pour le cas limite $k = K - 1$:

$$R_i^{\text{arr_K}}(\psi_{i,k}) = \mathbb{1}_{\text{term}_K(\psi_{i,K})}\ Cost_i^{\text{corr}}(\psi_{i,k}, a_i)$$

4.1.5 Fonction objectif (f)

La fonction objectif globale est définie comme la somme de tous les coûts pour l'ensemble des requêtes et des créneaux.

4.2 Formulation de DaCSA-PRMRP

Nous adoptons un schéma *Divide-and-Coordinate* [28] dans lequel chaque agent $A_{i,k}$ résout un problème d'optimisation local construit à partir du graphe de relations du DCOP global, ce qui conduit à notre méthode dédiée DaCSA-PRMRP.

Divide-and-Coordinate (DaC) Au-delà des méthodes de recherche locale et tabulaires, les approches *divide-and-coordinate* traitent la complexité des DCOP en *divisant* le DCOP global et en *coordonnant* itérativement les décisions locales [28]. Dans DaC, un DCOP binaire est décomposé en $|\mathcal{X}|$ sous-problèmes, où chaque sous-problème ϕ_s est défini sur x_s (contrôlé par l'agent s) ainsi que sur des copies de ses variables voisines dans le graphe de facteurs ; il inclut la relation unaire de x_s et la moitié de chaque relation binaire partagée avec ses voisins. Chaque sous-problème ϕ_s est formalisé comme un Programme Linéaire en Nombres Entiers (PLNE) binaire impliquant des variables $x_{i;k}^s$ valant 1 lorsque la variable x_i prend la valeur k, et des variables $x_{ij;kl}^s$ valant 1 lorsque x_i et x_j prennent respectivement les

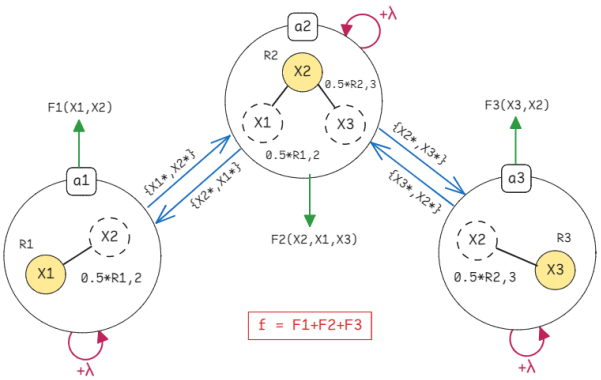


FIGURE 2 – Schéma de l'approche Divide-and-Coordinate.

valeurs k et l. Les contraintes suivantes doivent être vérifiées pour chaque ϕ_s : (i) des *contraintes d'unicité*, garantissant que chaque copie locale d'une variable reçoit exactement une valeur de son domaine ; (ii) des *contraintes de cohérence intra-sous-problème*, assurant la cohérence entre les représentations unaires et binaires d'une même variable au sein d'un même sous-problème ; (iii) des *contraintes de cohérence inter-sous-problèmes*, imposant l'accord entre copies locales de variables partagées et paires de variables partagées entre sous-problèmes voisins. Comme la troisième contrainte est globale, elle est relâchée par l'introduction de multiplicateurs de Lagrange dans l'objectif, sous la forme $\lambda^{j}_{i;k} \cdot (x^{i}_{i;k} - x^{j}_{i;k}) \quad \forall x_i \in \mathcal{X}, \forall x_j \in N(x_i), \forall k \in \mathcal{D}_i$ et $\lambda_{ij;kl} \cdot (x^{i}_{ij;kl} - x^{j}_{ij;kl}) \quad \forall r_{ij} \in \mathcal{R}, \forall k \in \mathcal{D}_i, \forall l \in \mathcal{D}_j$. Ainsi, chaque fonction objectif locale est réécrite pour intégrer ces termes lagrangiens, produisant un ensemble modifié de sous-problèmes dont les objectifs se somment en l'objectif initial du DCOP, constituant une *division valide* au sens où l'utilité globale est exactement retrouvée en agrégeant les utilités locales. Lors de l'étape de coordination, les agents échangent leurs solutions locales avec leurs voisins et mettent à jour les multiplicateurs de Lagrange en cas de désaccord ; ces multiplicateurs agissent comme des signaux de coordination pénalisant localement les affectations inconsistantes et guidant les agents vers un consensus.

4.2.1 Division du DCOP global

L'objectif du DCOP global est la somme des relations unaires (attachées à un seul créneau k) et des relations binaires couplant deux créneaux consécutifs $(k, k{+}1)$ d'une même requête i. Afin d'obtenir des sous-problèmes locaux indépendants tout en préservant l'objectif global, chaque agent $A_{i,k}$ définit un problème local $\phi_{i,k}$ formulé comme un Programme Linéaire en Nombres entiers (PLNE) binaire. Les variables utilisées ont déjà été décrites dans la section 4.2 ; on adapte les notations à notre problème :

— $x^{(i,k)}_{\mathrm{k};\psi}$,

— $x^{(i,k)}_{\mathrm{k-1};\psi'}$ $(k > 1)$, $x^{(i,k)}_{\mathrm{k+1};\psi''}$ $(k < K)$

— $x^{(i,k)}_{\mathrm{k-1,k};\psi',\psi}$ $(k > 1)$, $x^{(i,k)}_{\mathrm{k,k+1};\psi,\psi''}$ $(k < K)$

avec $\psi \in \mathcal{D}_{i,k}, \psi' \in \mathcal{D}_{i,k-1}, \psi'' \in \mathcal{D}_{i,k+1}$

Les relations du DCOP correspondant à des contraintes dures ne sont pas intégrées à la fonction objectif du sous-problème, mais traitées comme des contraintes du PLNE. La valeur de l'objectif repose ainsi exclusivement sur les relations de coût unaires et binaires, distinguées en deux ensembles :

(i) Toutes les relations unaires pour le créneau (i, k), définie par :

$$\mathcal{R}^{\mathrm{un}}_{i,k} = \left\{\mathcal{R}^{tronçon}, \mathcal{R}^{dep}, \mathcal{R}^{arr_K}\right\}$$

(ii) Toutes les relations binaires couplant les créneaux (i, k) et (i, j). On note cet ensemble :

$$\mathcal{R}^{\mathrm{bin}}_{i,k} = \left\{\mathcal{R}^{corr}, \mathcal{R}^{arr}\right\}$$

L'objectif local de l'agent $A_{i,k}$ est :

$$F_{i,k} = \sum_{R \in \mathcal{R}^{\mathrm{un}}} R(\psi_{i,k}) + \frac{1}{2} \sum_{j \in \{k-1, k+1\}} \sum_{R \in \mathcal{R}^{\mathrm{bin}}} R(\hat{\psi}_{i,j}, \psi_{i,k})$$

La relaxation de consensus dans DaCSA-PRMRP est obtenue en introduisant six familles de multiplicateurs de Lagrange imposant l'accord entre une variable globale et ses copies dans les sous-problèmes voisins : deux ensembles unaires par voisin ($\lambda^{\{k-1\}\to k}_{i,k}$ et $\lambda^{\{k+1\}\to k}_{i,k}$) et deux ensembles pour la direction inverse ($\lambda^{k\to\{k-1\}}_{i,k}$ et $\lambda^{k\to\{k+1\}}_{i,k}$), ainsi que deux ensembles par paires ($\lambda^{\{k-1\},k}_{i,k}$ et $\lambda^{k,\{k+1\}}_{i,k}$) gérant l'interaction entre copies adjacentes. Ces multiplicateurs définissent un terme lagrangien local $\mathcal{L}_{i,k}(\mathbf{x}, \lambda)$ qui pénalise tout désaccord entre les affectations choisies par des sous-problèmes voisins, à l'aide de fonctions indicatrices $\mathbf{1}[\cdot]$ pour activer les termes appropriés aux frontières de la chaîne.

$$\begin{aligned}
\mathcal{L}_{i,k}(\mathbf{x}, \lambda) = \sum_{\mathbb{D}^k_{i,k}} \Big(&\mathbb{I}[k > 1] \lambda^{\{k-1\}\to k}_{i,k} + \mathbb{I}[k < K] \lambda^{\{k+1\}\to k}_{i,k} \Big) x^{i,k}_{k} \\
&- \mathbb{I}[k > 1] \sum_{\mathbb{D}^{\{k-1\}}_{i,k}} \lambda^{k\to\{k-1\}}_{i,k} \, x^{i,k}_{\{k-1\}} \\
&- \mathbb{I}[k < K] \sum_{\mathbb{D}^{\{k+1\}}_{i,k}} \lambda^{k\to\{k+1\}}_{i,k} \, x^{i,k}_{\{k+1\}} \\
&+ \mathbb{I}[k > 1] \sum_{\mathbb{D}^{\{k-1\}}_{i,k} \times \mathbb{D}^{k}_{i,k}} \lambda^{\{k-1\},k}_{i,k} \, x^{i,k}_{\{k-1\},k} \\
&- \mathbb{I}[k < K] \sum_{\mathbb{D}^{k}_{i,k} \times \mathbb{D}^{\{k+1\}}_{i,k}} \lambda^{k,\{k+1\}}_{i,k} \, x^{i,k}_{k,\{k+1\}}
\end{aligned}$$

En combinant ce terme de consensus avec le coût local $F_{i,k}$, on obtient l'objectif local augmenté que chaque agent $A_{i,k}$ minimise à chaque itération : $\min \; F_{i,k} + \mathcal{L}_{i,k}$. Les détails des multiplicateurs sont fournis dans le matériel supplémentaire.

4.2.2 Mise à jour des lagrangiens

À chaque itération t, l'agent $A_{i,k}$ met à jour ses multiplicateurs à partir des solutions locales optimales des sous-problèmes adjacents.

Let $x^{\star,(i,k)}_{\cdot}$ denote the optimal solution of the subproblem (i,k) at iteration t. Soit $x^{\star,t}_{\cdot}$ la solution optimale du sous-problème considéré à l'itération t. Soit $\lambda^{\cdot}$ le multiplicateur de Lagrange à l'itération t et $\lambda^{\cdot,+}$ le même multiplicateur de Lagrange à l'itération suivante $t+1$. Les mises à jour s'écrivent :

$$\lambda_{i,k}^{\{k-1\}\to k,t+1} = \lambda_{i,k}^{\{k-1\}\to k,t} - \gamma_t\big(x_k^{\star,t} - x_{k-1}^{\star,t}\big),$$
$$\lambda_{i,k}^{\{k+1\}\to k,t+1} = \lambda_{i,k}^{\{k+1\}\to k,t} - \gamma_t\big(x_k^{\star,t} - x_{k+1}^{\star,t}\big),$$
$$\lambda_{i,k}^{k\to\{k-1\},t+1} = \lambda_{i,k}^{k\to\{k-1\},t} - \gamma_t\big(x_{k-1,k}^{\star,t} - x_{k-1}^{\star,t}\big),$$
$$\lambda_{i,k}^{k\to\{k+1\},t+1} = \lambda_{i,k}^{k\to\{k+1\},t} - \gamma_t\big(x_{k,k+1}^{\star,t} - x_{k+1}^{\star,t}\big)$$

Les multiplicateurs ne sont mis à jour qu'en cas de désaccord sur une variable partagée entre sous-problèmes voisins, l'amplitude de l'ajustement étant contrôlée par le pas $\gamma > 0$.

5 Résultats expérimentaux

5.1 Protocole expérimental

Environnement de calcul Les expériences ont été menées sur une station Linux équipée d'un processeur Intel Xeon E5-2660 v3 (20 cœurs à 2,60 GHz) et de 64 Go de RAM. Les algorithmes ont été implémentés en Python 3.9 à l'aide d'un fork du framework *pyDCOP*[1] pour la résolution DPOP et du framework *AMAK*[2] pour implémenter le modèle DaC multi-agents. Les MILP sont résolus avec IBM ILOG CPLEX Optimizer v22.11.

Instances du problème Afin d'évaluer la scalabilité et la qualité des solutions, nous avons généré des instances PRMRP synthétiques à partir de données de mobilité réelles fournies par un opérateur de transport public. Chaque instance du problème correspond à une requête de transport unique sur un réseau de transport public composé de lignes de bus et de métro à horaires fixes. Les réseaux utilisés incluent 5, 10 ou 15 lignes de transport, comprenant chacune 3 à 20 stations. Cela correspond à des réseaux allant de 15 à 300 nœuds de correspondance et à des tailles d'instance allant de 28 tronçons instanciés par créneau à 3092 tronçons instanciés par créneau. La Figure 3 présente le plus grand réseau utilisé pour ces expérimentations. Les données horaires réelles sont réduites jusqu'à 98% et toutes les expériences sont répliquées avec six graines aléatoires. Avant résolution, une étape de filtrage restreint les réseaux aux nœuds atteignables compte tenu des contraintes de départ, d'arrivée et la fenêtre temporelle de chaque requête. Par conséquent, la taille effective du problème dépend à la fois du réseau et de la requête. En moyenne, ce filtrage réduit le nombre de tronçons par créneau de 81,87%, avec des réductions allant de 70,25% à 100%.

Baselines Nous comparons DaCSA-PRMRP à une heuristique *gloutonne* centralisée inspirée de A*. Pour chaque requête, un itinéraire est construit incrémentalement en sélectionnant le meilleur tronçon suivant, minimisant son coût et une estimation du coût restant jusqu'à la destination. Cette stratégie rapide mais myope fournit une baseline (borne inférieure) sur la qualité des solutions. Nous comparons également les méthodes distribuées à une *formulation MILP centralisée*, implémentée avec CPLEX, qui fournit des solutions optimales et constitue une référence exacte pour la qualité.

1. https ://github.com/rpgoldman/pyDcop.git
2. https ://github.com/alexandreprl/AMAK-Python.git

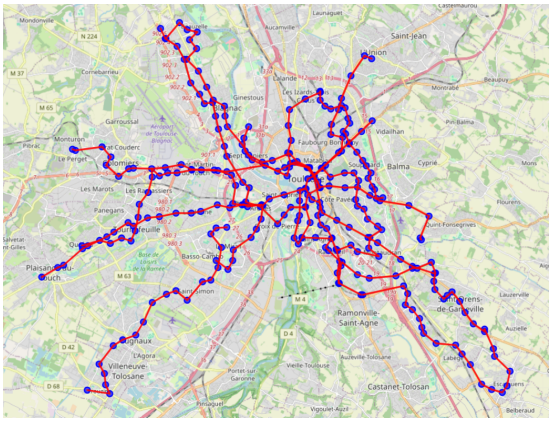


FIGURE 3 – Plus grand réseau de transport : 15 lignes, 20 arrêts par ligne.

Paramètres des modèles et des algorithmes Pour nos expériences, le nombre maximal de créneaux d'un itinéraire est fixé à $K = 5$ et le pas de discrétisation temporelle (pour obtenir les horaires de départ des ressources libres) est $\Delta = 5$ minutes. Le temps de résolution est limité à 5 minutes. Le timeout des sous-problèmes MILP dans DaCSA-PRMRP est fixé à 30 secondes. DaCSA-PRMRP s'arrête lorsque l'optimalité est atteinte avec un gap d'optimalité inférieur à $\epsilon = 0.0001$ ou lorsque la solution n'a pas été améliorée depuis 4 itérations.
La règle de mise à jour du pas dans DaCSA-PRMRP suit une règle de sous-gradient adaptatif proposée dans [28]. Pour des raisons de stabilité, le pas a été borné dans l'intervalle $[0.7, 20]$, ces bornes ayant été choisies empiriquement.

5.2 Analyse des résultats

Performances Notre analyse évalue la scalabilité en fonction de la taille des instances (nombre de tronçons avant filtrage) via le temps de résolution et la charge de calcul. Des tendances lissées (moyennes mobiles) sont utilisées afin de mettre en évidence les motifs de croissance tout en réduisant la variabilité inter-instances. La Figure 4 présente l'évolution du temps de résolution. En tant qu'algorithme anytime, MGM-2 atteint systématiquement la limite de temps. L'heuristique gloutonne surpasse sans surprise les autres méthodes, tandis que CPLEX reste applicable sur des instances plus grandes. Les temps de résolution de DPOP et de DaCSA-PRMRP croissent fortement et présentent des performances moyennes comparables. Ils commencent à atteindre la limite de temps respectivement autour de 1254 tronçons instanciés et 730 pour DaCSA-PRMRP. Contrairement à DPOP, DaCSA-PRMRP garantit une solution de

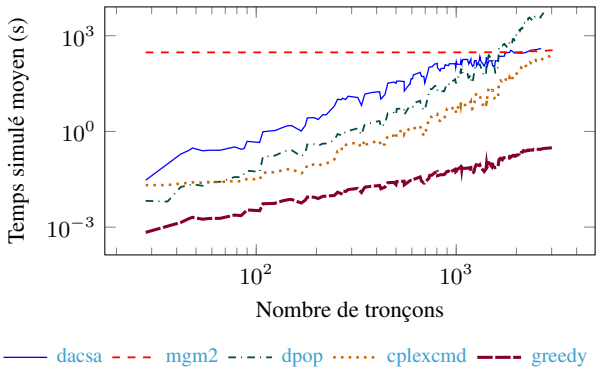


FIGURE 4 – Temps simulé moyen en secondes (axe y en log) en fonction du nombre de tronçons (axe x en log).

qualité bornée après la limite de temps.

Concernant la charge de communication, la 5-(a) montre que MGM-2 induit un surcoût de communication très élevé, tandis que DPOP échange un nombre de messages presque constant. DaCSA-PRMRP présente une complexité en messages linéaire avec le nombre d'itérations, ce qui se traduit par une variabilité visible d'une instance à l'autre, tout en conservant un comportement globalement stable. La diminution du nombre de messages observée pour certaines méthodes, en particulier MGM-2, est principalement due à des mécanismes d'arrêt anticipé qui limitent le nombre d'itérations lorsque la taille des instances augmente et que les itérations deviennent plus coûteuses. Cet effet peut également être renforcé par des propriétés structurelles des grands réseaux de transport public, qui comportent davantage de lignes traversantes et nécessitent donc moins de coordination lors des correspondances.

La 5-(b) rapporte l'évolution de la taille totale des messages. Comme attendu pour une méthode tabulaire dont le coût de communication croît avec la dimension des tables, la taille des messages de DPOP augmente rapidement avec la taille des instances. MGM-2 présente d'abord une diminution liée à l'arrêt anticipé, puis une augmentation lorsque la taille unitaire des messages devient dominante. DaCSA-PRMRP présente un comportement globalement stable, avec une légère baisse sur les plus grandes instances, probablement due au timeout ou à des effets structurels.

Qualité des solutions Au-delà du temps de calcul et de la scalabilité, nous analysons la qualité des solutions. Les violations de contraintes sont rares ($\leq 7.94\%$ des instances) : CPLEX et DPOP violent des contraintes dans 0,62% des cas, MGM-2 dans 3,70%, et DaCSA-PRMRP dans 7,94%. Les violations concernent toujours une seule contrainte parmi les 14 possibles (issues des contraintes dures DCOP) et portent sur les temps minimaux de correspondance dans 78% des instances affectées, les autres cas correspondant à des violations de la règle interdisant les solutions « marche uniquement ».

Nous avons observé l'écart relatif moyen à CPLEX en fonction de la taille des instances. Tous les algorithmes sont proches de l'optimum sur les petits graphes, et l'écart relatif moyen atteint son maximum autour de 2% pour les plus grandes instances.

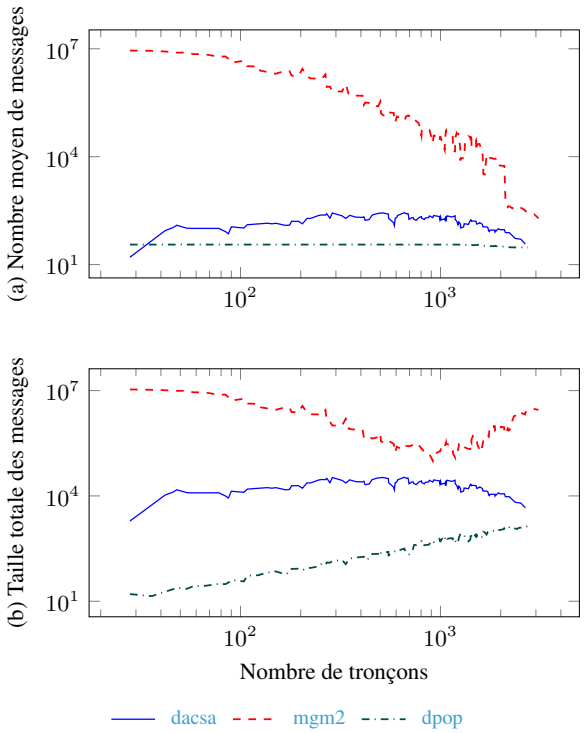


FIGURE 5 – Charge de communication en fonction du nombre de tronçons (log-log) : (a) nombre moyen de messages ; (b) taille totale moyenne des messages.

6 Conclusion

Dans ce travail, nous avons introduit le PRMRP, un modèle formel capturant la complexité complète de la planification d'itinéraires multimodaux et multi-utilisateurs dans un contexte MaaS. À partir de cette formulation, nous avons dérivé deux représentations distribuées complémentaires (DCOP et DaCSA-PRMRP). Des expériences menées sur des instances synthétiques dérivées de données de mobilité réelles montrent que les deux approches atteignent une bonne qualité de solution. De plus, DaCSA-PRMRP présente une meilleure scalabilité : il nécessite moins de messages que MGM-2 et, contrairement à DPOP, il garantit une solution de qualité bornée, ce qui lui permet de traiter des réseaux plus grands et des fenêtres temporelles plus contraintes. Ces résultats confirment qu'une formulation distribuée, respectueuse de la confidentialité, est non seulement faisable mais aussi avantageuse pour des systèmes MaaS à grande échelle où un contrôle centralisé est impraticable.

L'expérimentation réalisée s'est concentrée sur des instances à deux modes de transport (bus/métro) avec des ressources fixes. Néanmoins, notre formulation du problème prend en compte des ressources libres ainsi que les préférences des utilisateurs concernant leurs ressources de mobilité personnelles. L'inclusion de ressources libres dans

les expériences entraînerait un plus grand nombre de combinaisons multimodales et davantage de personnalisation. Nous prévoyons d'évaluer cela à l'aide de trois scénarios de mobilité combinant usage de voiture privée et transport public, afin de mettre en évidence la pertinence du report modal : (1) sans voiture, (2) voiture privée avec stationnement à destination, et (3) voiture privée avec stationnements situés à certains points d'entrée des lignes de transport public. En complément, une optimisation selon plusieurs critères (durée, émissions de CO_2, prix par exemple) sous différents profils de préférences (orienté temps, orienté CO_2, orienté coût) pourrait mettre en évidence des interactions significatives entre multimodalité et préférences utilisateurs, conduisant à des arbitrages modaux distincts selon les scénarios.

L'étude actuelle repose sur plusieurs hypothèses simplificatrices, telles que l'absence de contraintes explicites de capacité des ressources et l'indépendance des requêtes. Les travaux futurs porteront donc sur : (i) l'intégration du partage des ressources (p. ex. capacité des véhicules, limites des stations d'arrimage) directement dans les modèles DCOP/DaC, possiblement via des contraintes de couplage supplémentaires ou des couches de coordination hiérarchiques ; (ii) des stratégies hybrides combinant les heuristiques rapides de la baseline gloutonne avec l'exactitude de DaC, visant des algorithmes anytime capables de fournir des solutions progressivement meilleures sous des budgets temps stricts.

Globalement, le cadre d'optimisation distribuée proposé ouvre une voie prometteuse vers un MaaS scalable et centré utilisateur, capable de respecter conjointement les préférences personnelles, les contraintes opérationnelles et la confidentialité des agents participants.

Références

[1] Parastoo Afrasyabi, Mohammad Saadi Mesgari, El sayed M. El-kenawy, Mehrdad Kaveh, Abdelhameed Ibrahim, and Nima Khodadadi. A crossover-based multi-objective discrete particle swarm optimization model for solving multi-modal routing problems. *Decision Analytics Journal*, 9 :100356, 2023.

[2] Chris Barrett, Riko Jacob, and Madhav Marathe. Formal-language-constrained path problems. *SIAM J. Comput.*, 30 :809–837, 01 2000.

[3] Hannah Bast. Car or public transport–two worlds. 5760 :355–367, 09 2009.

[4] Hannah Bast, Daniel Delling, Andrew Goldberg, Matthias Müller-Hannemann, Thomas Pajor, Peter Sanders, Dorothea Wagner, and Renato F. Werneck. Route planning in transportation networks, 2015.

[5] Jean-François Cordeau. A branch-and-cut algorithm for the dial-a-ride problem. *Operations Research*, 54 :573–586, 06 2006.

[6] Daniel Delling, Thomas Pajor, and Dorothea Wagner. *Engineering Time-Expanded Graphs for Faster Timetable Information*, pages 182–206. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[7] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1) :269–271, December 1959.

[8] Mariagrazia Dotoli, Hayfa Zgaya, Carmine Russo, and Slim Hammadi. A multi-agent advanced traveler information system for optimal trip planning in a co-modal framework. *IEEE Transactions on Intelligent Transportation Systems*, 18(9) :2397–2412, 2017.

[9] European Parliament and Council of the EU. Regulation (EU) 2016/679 of the European Parliament and of the Council. https://data.europa.eu/eli/reg/2016/679/oj, 2016.

[10] A. Farinelli, A. Rogers, A. Petcu, and N. R. Jennings. Decentralised coordination of low-power embedded devices using the max-sum algorithm. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '08, page 639–646, Richland, SC, 2008. International Foundation for Autonomous Agents and Multiagent Systems.

[11] Ferdinando Fioretto, Enrico Pontelli, and William Yeoh. Distributed constraint optimization problems and applications : A survey. *Journal of Artificial Intelligence Research*, 61 :623–698, 2018.

[12] K. Gkiotsalitis and A. Nikolopoulou. The multi-vehicle dial-a-ride problem with interchange and perceived passenger travel times. *Transportation Research Part C : Emerging Technologies*, 156 :104353, 2023.

[13] Katsutoshi Hirayama and Makoto Yokoo. Distributed partial constraint satisfaction problem. In Gert Smolka, editor, *Principles and Practice of Constraint Programming-CP97*, pages 222–236, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg.

[14] Alexandra Hoess, Jonathan Lautenschlager, Johannes Sedlmeir, Gilbert Fridgen, Vincent Schlatt, and Nils Urbach. Toward seamless mobility-as-a-service. *Business & Information Systems Engineering*, 67(2) :213–229, 2025.

[15] Carl H. Häll, Henrik Andersson, Jan T. Lundgren, and Peter Värbrand. The Integrated Dial-a-Ride Problem. *Public Transport*, 1(1) :39–54, May 2009.

[16] Peraphan Jittrapirom, Valeria Caiati, Anna Maria Feneri, Shima Ebrahimigharehbaghi, Maria Alonso Gonzalez, and Jishnu Narayan. Mobility as a service : A critical review of definitions, assessments of schemes, and key challenges. *Urban Planning*, 2, 06 2017.

[17] Pramesh Kumar and Alireza Khani. An algorithm for integrating peer-to-peer ridesharing and schedule-based transit system for first mile/last mile access. *Transportation Research Part C : Emerging Technologies*, 122 :102891, 01 2021.

[18] Tai-Yu Ma, Saeid Rasulkhani, Joseph Y.J. Chow, and Sylvain Klein. A dynamic ridesharing dispatch

and idle vehicle repositioning strategy with integrated transit transfers. *Transportation Research Part E : Logistics and Transportation Review*, 128 :417–442, 2019.

[19] Rajiv Maheswaran, Jonathan Pearce, and Milind Tambe. Distributed algorithms for dcop : A graphical-game-based approach. pages 432–439, 01 2004.

[20] Neda Masoud and R. Jayakrishnan. A decomposition algorithm to solve the multi-hop peer-to-peer ride-matching problem. *Transportation Research Part B : Methodological*, 99 :1–29, 2017.

[21] Renaud Masson, Fabien Lehuédé, and Olivier Péton. The dial-a-ride problem with transfers. *Computers & Operations Research*, 41 :12–23, 2014.

[22] Pragnesh Jay Modi, Wei-Min Shen, Milind Tambe, and Makoto Yokoo. Adopt : asynchronous distributed constraint optimization with quality guarantees. *Artificial Intelligence*, 161(1) :149–180, 2005. Distributed Constraint Satisfaction.

[23] Sophie N. Parragh. Introducing heterogeneous users and vehicles into models and algorithms for the dial-a-ride problem. *Transportation Research Part C : Emerging Technologies*, 19(5) :912–930, 2011. Freight Transportation and Logistics (selected papers from ODYSSEUS 2009 - the 4th International Workshop on Freight Transportation and Logistics).

[24] Adrian Petcu and Boi Faltings. A scalable method for multiagent constraint optimization. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, page 266–271, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[25] Marcus Posada, Henrik Andersson, and Carl Häll. The integrated dial-a-ride problem with timetabled fixed route service. *Public Transport*, 9, 07 2017.

[26] Tamir Tassa, Tal Grinshpoun, and Avishay Yanai. A privacy preserving collusion secure dcop algorithm. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 4774–4780. AAAI Press, 2019.

[27] Daniel Tischner. Multi-modal route planning in road and transit networks, 2018.

[28] Meritxell Vinyals, Marc Pujol, Juan A. Rodríguez-Aguilar, and Jesús Cerquides. Divide and coordinate : solving dcops by agreement. In *Adaptive Agents and Multi-Agent Systems*, 2010.

[29] Shuang Yang, Jianjun Wu, Huijun Sun, and Yunchao Qu. Trip planning for a mobility-as-a-service system : Integrating metros and shared autonomous vehicles. *Transportation Research Part E : Logistics and Transportation Review*, 176 :103217, 2023.

[30] Xiao Yu, Huimin Miao, Armagan Bayram, Meigui Yu, and Xi Chen. Optimal routing of multimodal mobility systems with ride-sharing. *International Transactions in Operational Research*, 28(3) :1164–1189, 2021.

[31] Yimeng Zhang, Oded Cats, and Shadi Sharif Azadeh. Dynamic preference-based multi-modal trip planning of public transport and shared mobility, 2025.

SESSION 7 : IA et Humains présidée par Nahla Ben Amor

Approche hybride centrée sur l'humain pour l'évaluation explicable de la qualité de présentation des tweets organisationnels

Aya Hidri[1], Ines Saad[2, 3], Manel Ben Sassi[1], Henda Ben Ghezala[1]

[1] Laboratoire RIADI, Université de la Manouba
[2]Laboratoire MIS, Université de Picardie Jules Verne
[3]Amiens Business School

Résumé

Cet article propose un cadre hybride multicritère centré sur le concept de Human-In-The-Loop (HITL) pour l'évaluation explicable de la qualité de présentation textuelle des tweets organisationnels. L'approche combine des métriques de lisibilité classiques (Flesch Reading Ease, Kandel-Moles), une évaluation réalisée par le grand modèle de langage Llama-3.3-70B, et l'expertise humaine, agrégées via la méthode de la somme pondérée (SAW). Les résultats montrent que Llama-3.3-70B atteint une corrélation de Pearson de $r = 0{,}52$ avec les jugements de l'expert, surpassant les métriques classiques (FRE : $r = 0{,}17$; KM : $r = 0{,}21$). Ces résultats confirment l'intérêt d'une approche hybride intégrant l'intelligence artificielle et l'expertise humaine pour l'évaluation contextualisée de la qualité de présentation textuelle des tweets organisationnels.

Mots-clés

Qualité des données ; X (Twitter) ; Human-In-The-Loop ; grands modèles de langage (LLM) ; lisibilité textuelle.

Abstract

This paper proposes a hybrid multicriteria Human-In-The-Loop (HITL) approach for the explainable evaluation of the textual presentation quality of organizational tweets. The approach combines classical readability metrics (Flesch Reading Ease, Kandel-Moles), an evaluation by the large language model Llama-3.3-70B, and human expertise, aggregated via the Simple Additive Weighting (SAW) method. Our results show that Llama-3.3-70B achieves a Pearson correlation of $r = 0.52$ with expert judgments, outperforming classical metrics (FRE : $r = 0.17$; KM : $r = 0.21$). These results confirm the relevance of a hybrid approach combining artificial intelligence and human expertise for the contextualized assessment of the textual presentation quality of organizational tweets.

Keywords

Data quality ; X (Twitter) ; Human-In-The-Loop ; large language models (LLMs) ; text readability.

1 Introduction

X (anciennement Twitter) a révolutionné la communication numérique, permettant aux utilisateurs d'exprimer leurs opinions, de suivre l'actualité et de partager du contenu avec une audience mondiale. Avec plus de 440 millions d'utilisateurs actifs mensuels en 2024, cette plateforme génère quotidiennement près de 500 millions de tweets [2]. Cette quantité massive de données, souvent qualifiée de *big data*, a suscité une attention considérable dans la recherche académique [24].

Les entreprises et les organisations exploitent stratégiquement ces données pour surveiller leur réputation, évaluer la perception de leur marque et adapter leurs stratégies de communication[22]. Des entreprises telles que Nike ou Coca-Cola cherchent à s'assurer que leur positionnement est aligné avec les valeurs sociales actuelles et les attentes des consommateurs [17]. Ces entreprises doivent donc comprendre comment leurs parties prenantes réagissent aux tendances émergentes pertinentes pour leur secteur.

X intègre différents formats de contenu multimodaux (texte, images, vidéos et audio) afin d'améliorer l'engagement des utilisateurs et la visibilité des publications. Plusieurs travaux ont démontré que l'utilisation d'éléments visuels contribue à renforcer l'attractivité des messages et à stimuler l'interaction[18][14]. Toutefois, au-delà de ces éléments multimédias, la qualité textuelle demeure un facteur déterminant dans l'efficacité des communications organisationnelles sur les réseaux sociaux[6].

Cependant, la nature intrinsèque des tweets, souvent caractérisée par l'utilisation d'abréviations, d'emojis et de hashtags, pose des défis majeurs pour l'analyse sémantique. Une information mal structurée peut nuire à la compréhension du message, et limiter son utilité dans le cadre de la prise de décision organisationnelle.

Ces constats soulèvent une problématique centrale : comment évaluer de manière fiable, explicable et adaptée au contexte organisationnel la qualité de présentation textuelle des tweets, tout en tenant compte de leur brièveté et de leur style informel ?

C'est pourquoi nous proposons une approche hybride combinant l'automatisation et l'expertise humaine au sein d'un cadre multicritère, permettant de produire un score global explicable de qualité présentation textuelle.

Les questions de recherche suivantes guident cette étude :

QR1 : Dans le contexte des tweets organisationnels, dans quelle mesure les scores de lisibilité produits par Llama-

3.3-70B corrèlent-ils avec les jugements d'expert, comparativement aux métriques traditionnelles (Flesch Reading Ease, Kandel-Moles) ?
QR2 : Existe-t-il des relations significatives entre la lisibilité du texte et d'autres caractéristiques textuelles propres à la plateforme X (hashtags, emojis) ?
QR3 : Quel est l'apport d'une approche hybride intégrant l'expertise humaine dans l'évaluation de la qualité de présentation textuelle des tweets pour la prise de décision organisationnelle ?
Ces trois questions permettent d'adresser un double enjeu : d'une part, comprendre l'influence des éléments spécifiques à la plateforme X sur la lisibilité perçue ; d'autre part, évaluer la valeur ajoutée d'un cadre multicritère intégrant l'expertise humaine face aux approches automatiques. La réponse à ces questions apporte une contribution méthodologique originale à l'intersection de la qualité des données, de l'analyse des réseaux sociaux et des systèmes d'aide à la décision.
Le reste de cet article est organisé comme suit. La section 2 présente les travaux antérieurs. La section 3 introduit le cadre formel d'évaluation. La section 4 décrit les méthodes d'évaluation de la lisibilité. La section 5 présente les résultats et leur analyse. Enfin, la section 6 expose la conclusion et les perspectives de recherche futures.

2 Travaux antérieurs

L'analyse des contenus publiés sur les réseaux sociaux s'articule autour de trois axes principaux dans la littérature : l'impact des éléments visuels et textuels sur l'engagement, l'évaluation de la lisibilité des contenus textuels, et les approches automatiques fondées sur l'apprentissage automatique et les LLM. Ces trois axes constituent le socle théorique sur lequel repose notre contribution.

2.1 Éléments visuels, hashtags et emojis

Plusieurs études ont mis en évidence l'impact des éléments visuels tels que les images et les vidéos sur l'engagement des utilisateurs sur les réseaux sociaux[18][6]. Ces éléments sont traités plus rapidement que le texte, ce qui leur confère un avantage considérable en matière d'attractivité et d'impact émotionnel. Ils jouent également un rôle déterminant dans les décisions d'achat et dans l'élaboration des stratégies de communication digitale. À cet égard, des travaux empiriques ont démontré que les publications intégrant des images ou des vidéos génèrent significativement plus d'engagement positif de la part des utilisateurs [14]. Des recherches plus récentes confirment par ailleurs que la combinaison d'un texte bien rédigé et d'éléments visuels pertinents permet d'accroître la portée et la mémorabilité des messages organisationnels [1].
Au-delà des éléments visuels, d'autres travaux se sont intéressés aux caractéristiques textuelles spécifiques aux publications sur les réseaux sociaux, notamment à l'utilisation des hashtags et des emojis [25, 8]. Ces éléments contribuent à accroître la visibilité des publications, à augmenter le nombre d'impressions et à renforcer la notoriété de la marque. Les auteurs de [8] montrent que les hashtags servent principalement à augmenter la visibilité des publications et à renforcer l'identification des organisations sur les plateformes de réseaux sociaux. Par ailleurs, les auteurs de [11][15] ont mis en évidence que l'usage approprié des hashtags et des emojis permet de renforcer le ton émotionnel du message et d'améliorer la connexion avec l'audience, à condition toutefois de ne pas surcharger le texte au détriment de sa lisibilité.

2.2 Lisibilité des contenus numériques courts

Plus récemment, la lisibilité des contenus numériques a émergé comme un facteur déterminant de leur valeur perçue et de leur attractivité. Elle est définie dans la littérature comme la facilité avec laquelle un texte peut être compris par les lecteurs[19]. La lisibilité est ainsi décrite comme la « facilité de compréhension due au style d'écriture », ou encore comme le degré auquel un texte peut être appréhendé par les lecteurs[5].
La lisibilité dépend de plusieurs caractéristiques linguistiques couvrant différents niveaux du texte, notamment le vocabulaire utilisé, la longueur des mots, la complexité syntaxique, la longueur des phrases, ainsi que la cohésion textuelle [26]. Ces dimensions influencent directement la capacité des lecteurs à traiter efficacement l'information. Afin de mesurer la lisibilité de manière quantitative, plusieurs indices ont été proposés dans la littérature. Parmi les plus utilisés figurent l'indice de Flesch-Kincaid (FK) et le Gunning Fog Index (GFI) [10], qui estiment respectivement le niveau scolaire requis pour comprendre un texte en se basant sur la longueur moyenne des phrases et des mots. Ces indices visent à évaluer la difficulté de compréhension et à limiter les complexités inutiles dans la rédaction des contenus.
Cependant, la plupart des travaux sur la lisibilité se concentrent principalement sur les textes longs (contenus web ou pédagogiques), tandis que les contenus courts tels que les publications sur les réseaux sociaux restent peu étudiés [10]. Or, ces contenus présentent des spécificités importantes, notamment l'usage d'abréviations, de hashtags, d'emojis et de structures informelles, qui compliquent l'application des indices classiques [11]. Plus récemment, des travaux ont mis en évidence que certaines caractéristiques textuelles, notamment les hashtags et les mentions, sont négativement corrélées avec la lisibilité perçue des tweets [25]. Ces travaux suggèrent que la lisibilité ne constitue pas seulement un critère de qualité textuelle, mais également un levier stratégique pour les organisations dans leurs communications numériques.

2.3 Approches automatiques et grands modèles de langage

Face aux limitations des métriques classiques, des approches supervisées ont abordé l'évaluation de la lisibilité comme un problème de classification automatique, fondées sur les caractéristiques lexicales telles que la longueur des mots et la complexité lexicale[20][26]. Cependant, ces caractéristiques ne parviennent souvent pas à capturer le contexte des phrases. Certains travaux ont tenté

de pallier ces limitations en combinant des représentations n-grammes avec des méthodes d'apprentissage automatique notamment les Support Vector Machines (SVM), les Random Forest et les réseaux de neurones artificiels (ANN)[26][7]. L'application des méthodes d'apprentissage automatique à l'évaluation de la lisibilité textuelle présente un potentiel certain ; toutefois, ces approches requièrent des corpus d'entraînement annotés de grande envergure et présentent des limitations significatives en termes de généralisation interlangues et d'adaptation contextuelle. De surcroît, elles ne parviennent pas à capturer le contexte global ni la sémantique profonde du texte.

Les récentes avancées en traitement automatique du langage naturel (TALN) et en intelligence artificielle générative (IAG) ont ouvert de nouvelles perspectives pour l'évaluation automatique de la lisibilité des textes. Plusieurs travaux ont commencé à explorer le potentiel des grands modèles de langage (LLM) dans ce contexte [21][3][19][4]. Par exemple, les auteurs de [3] ont examiné l'applicabilité de ChatGPT en tant qu'outil d'évaluation de la lisibilité, en soumettant des manuels scolaires sélectionnés à une analyse textuelle par LLM. Cette étude démontre que les LLM constituent une alternative prometteuse aux formules traditionnelles, notamment pour les contextes pédagogiques spécifiques aux apprenants en langue étrangère. Dans le même contexte, les auteurs de [19] ont conduit une étude empirique analysant les capacités de génération textuelle de plusieurs LLM pré-entraînés, en adoptant la lisibilité comme métrique d'évaluation principale. Les résultats montrent que la lisibilité varie significativement selon le modèle utilisé et les contraintes imposées. Les auteurs soulignent toutefois que les métriques traditionnelles de lisibilité seules demeurent insuffisantes pour capturer l'ensemble des dimensions de la lisibilité, telles que la pertinence contextuelle et l'engagement du lecteur. Enfin, les auteurs de [4] ont démontré que les métriques traditionnelles, telles que le Flesch–Kincaid Grade Level (FKGL), présentent une faible corrélation avec les jugements humains. En revanche, les LLM montrent une meilleure corrélation avec les évaluations humaines. Ces résultats s'expliquent par la capacité des LLM à capturer des dimensions plus profondes de la lisibilité :le niveau de connaissances préalables requis pour comprendre un texte, la cohérence discursive et l'adéquation au lectorat cible. Ces aspects demeurent difficilement appréhendables par les formules syllabiques classiques.

Néanmoins, une limitation majeure subsiste : la plupart de ces méthodes ont été conçues pour des textes de longueur standard et ne sont pas directement adaptées aux contenus courts et informels des réseaux sociaux. De plus, peu de travaux proposent une approche intégrant explicitement les préférences des décideurs dans l'évaluation de la qualité textuelle des tweets organisationnels. C'est précisément cette double lacune tels que l'inadaptation des métriques classiques aux contenus courts et l'absence de prise en compte des préférences décisionnelles que notre travail cherche à combler en proposant un cadre hybride multicritère centré sur l'humain, combinant métriques classiques, évaluation par LLM (Llama-3.3-70B) et expertise humaine, offrant ainsi robustesse, explicabilité et alignement sur les besoins organisationnels. Dans ce contexte, cette étude vise à évaluer la qualité de présentation textuelle des tweets organisationnels. L'objectif est d'intégrer les préférences des décideurs afin de proposer une approche d'évaluation adaptée au contexte organisationnel. La méthodologie proposée repose sur une architecture hybride combinant l'expertise humaine via un paradigme HITL et un traitement algorithmique automatisé.

3 Évaluation de la qualité de présentation textuelle

Le critère de qualité de présentation textuelle a été consolidé par la validation d'un expert issu du domaine du marketing des médias sociaux et de la communication organisationnelle. Ce critère est évalué à travers sa dimension de lisibilité textuelle, ainsi que l'impact des hashtags et l'usage des emojis, qui constituent des éléments centraux du message dans les tweets. Il mesure la manière dont l'information est structurée et perçue par les différents publics ciblés : candidats potentiels, partenaires académiques et autres parties prenantes institutionnelles. L'approche proposée, illustrée par la Figure 1, s'articule en trois phases.

Phase 1 : Acquisition et préparation des données
Cette phase porte sur la constitution d'un corpus de 11 000 tweets issus des comptes officiels d'écoles de commerce françaises sur la plateforme X, avec extraction des éléments textuels, des hashtags, des emojis et des liens originaux.

Phase 2 : Evaluation hybride HITL
Cette phase combine une évaluation automatique de la lisibilité (FRE, Kandel-Moles et Llama-3.3-70B) et une intervention experte (HITL) consistant à sélectionner les tweets de référence via la méthode DRSA [9][23], à construire les sous-critères et à définir les poids associés.

Phase 3 : Validation et agrégation multicritère
Cette phase assure la validation par mise en correspondance des scores automatiques avec les jugements de l'expert sur les 50 tweets de référence, puis l'agrégation multicritère via la méthode SAW afin de produire un score explicable de la qualité de présentation textuelle.

Dans cette section, nous présentons le modèle mathématique multicritère fondant notre approche. La Section 3.1 établit les notations fondamentales. La Section 3.2 formalise la représentation d'un tweet organisationnel. La Section 3.3 présente la méthode de la somme pondérée (SAW).

3.1 Notations fondamentales

Soit $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$ l'ensemble des m tweets à évaluer, où T_i ($i = 1, 2, \ldots, m$) désigne un tweet individuel. Un tweet est une séquence de tokens :

$$T_i = \langle w_1, w_2, \ldots, w_k \rangle \tag{1}$$

où les tokens w_ℓ ($\ell = 1, 2, \ldots, k$) peuvent appartenir à l'une des trois catégories textuelles : le contenu textuel, les hashtags ou les emojis.

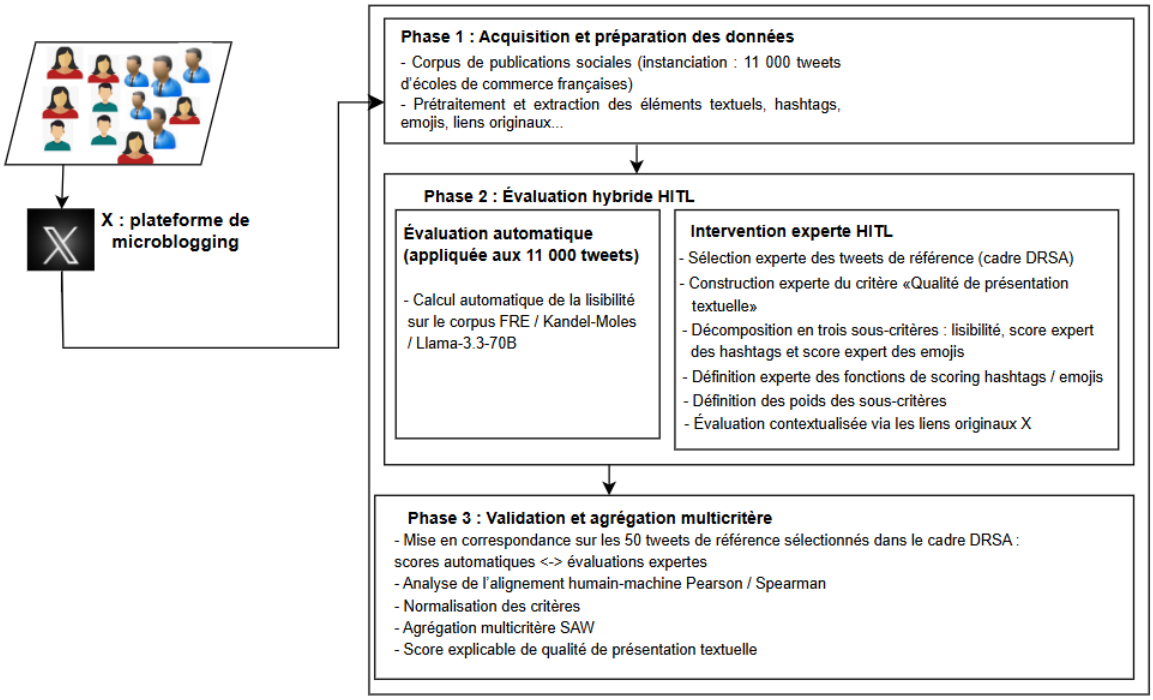


FIGURE 1 – Cadre hybride pour l'évaluation explicable de la qualité de présentation textuelle des tweets organisationnels.

Soit $\mathcal{G} = \{g_1, \ldots, g_n\}$ l'ensemble des n critères d'évaluation, où g_j $(j = 1, 2, \ldots, n)$ désigne le j-ème critère.

3.2 Représentation formelle d'un tweet

Un tweet $T_i \in \mathcal{T}$ est représenté par le triplet :

$$T_i \;=\; \big\langle\, \mathcal{C}(T_i),\; \mathcal{H}(T_i),\; \mathcal{E}(T_i) \,\big\rangle \tag{2}$$

où :

— $\mathcal{C}(T_i) \in \Sigma^*$ désigne la séquence ordonnée de tokens lexicaux porteurs du message principal, avec Σ^* l'ensemble de toutes les séquences finies de tokens ;

— $\mathcal{H}(T_i)$ désigne l'ensemble fini des hashtags (préfixés par #) assurant la catégorisation thématique et la découvrabilité du contenu ;

— $\mathcal{E}(T_i)$ désigne l'ensemble fini des emojis contribuant à l'expressivité non verbale et à la tonalité émotionnelle du message.

3.3 Agrégation SAW et score de présentation textuelle

Nous adoptons la méthode SAW [12] pour agréger les n critères en un score scalaire global. Cette méthode garantit la traçabilité de la contribution de chaque critère et l'interprétabilité directe du score par les décideurs.

Étape 1 : Construction de la matrice de décision
La matrice de décision initiale $X \in R^{m \times n}$ est définie comme suit :

$$X = (x_{ij}) = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \tag{3}$$

où chaque ligne i correspond au tweet $T_i \in \mathcal{T}$, chaque colonne j au critère $g_j \in \mathcal{G}$, et $x_{ij} = g_j(T_i)$.

Étape 2 : Normalisation de la matrice de décision
Afin de rendre les n critères comparables sur une échelle commune $[0, 1]$, on suppose $x_{ij} > 0$ pour tout $i \in \{1, \ldots, m\}$ et $j \in \{1, \ldots, n\}$, ce qui garantit que les formules de normalisation ci-dessous sont bien définies. Chaque valeur de performance brute x_{ij} est normalisée selon la nature du critère g_j :

$$\tilde{x}_{ij} = \begin{cases} \dfrac{x_{ij}}{\max_i x_{ij}}, & \text{si } g_j \text{ est un critère bénéfice,} \\[2ex] \dfrac{\min_i x_{ij}}{x_{ij}}, & \text{si } g_j \text{ est un critère coût.} \end{cases} \tag{4}$$

Un critère est dit *bénéfice* lorsque des valeurs plus élevées sont préférables, et *coût* lorsque des valeurs plus faibles sont préférées. Les 3 critères retenus étant tous de type bénéfice, la matrice normalisée est :

$$\tilde{X} = (\tilde{x}_{ij}) \;\in\; [0, 1]^{m \times n}, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, n \tag{5}$$

Étape 3 : Définition du vecteur de pondération
Soit $W = (W_1, \ldots, W_n)^\top \in [0, 1]^n$ le vecteur de pondération, où W_j traduit l'importance relative du critère g_j, établi lors des sessions avec l'expert.

Ce vecteur vérifie la contrainte de normalisation :

$$\sum_{j=1}^{n} W_j = 1, \qquad W_j \geq 0 \quad \forall\, j \in \{1, \ldots, n\} \tag{6}$$

Étape 4 : Calcul du score global
Le score global de présentation textuelle S_i du tweet $T_i \in \mathcal{T}$ est défini par la somme pondérée des performances normalisées :

$$S_i \;=\; W^{\top} \tilde{x}_i \;=\; \sum_{j=1}^{n} W_j \, \tilde{x}_{ij}, \qquad i = 1, \ldots, m \tag{7}$$

où $\tilde{x}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{in})^{\top} \in [0,1]^n$ est le vecteur des performances normalisées du tweet T_i, avec $S_i \in [0,1]$.

4 Méthodes d'évaluation de la lisibilité

Pour mesurer la lisibilité, deux métriques traditionnelles ont été employées : Flesch Reading Ease pour l'anglais et la formule de Kandel-Moles pour le français. Pour les tweets en anglais, l'indice Flesch Reading Ease mesure la difficulté en fonction de la longueur des phrases et de la densité syllabique[16]. Pour les tweets en français, la formule de Kandel-Moles est appliquée sur la base du nombre de mots et de syllabes[13]. Bien que les métriques classiques soient robustes, elles restent limitées aux caractéristiques de surface (longueur des mots et des phrases). Pour capturer la véritable difficulté linguistique, nous intégrons le modèle Llama-3.3-70B via l'API Groq.

4.1 Kandel–Moles pour le contenu en français

Pour évaluer la lisibilité des tweets en langue française, nous utilisons la formule de Kandel et Moles. Cette métrique est spécifiquement conçue pour la structure phonétique et syllabique du français, ce qui la rend particulièrement adaptée à l'analyse de contenus numériques en français. Contrairement aux indices Dale-Chall ou Gunning Fog, qui s'appuient sur de larges corpus de référence ou des listes de « mots difficiles » souvent peu fiables pour les contenus des réseaux sociaux en raison de l'argot et des abréviations, la formule de Kandel-Moles repose sur des mesures orthographiques et syllabiques.
La formule a été légèrement adaptée pour tenir compte du format court des tweets. Compte tenu de leur brièveté et de leur structure informelle, chaque tweet est traité comme une phrase fonctionnelle unique ($N_{\text{phrases}} = 1$). La formule est définie comme suit :

$$RE_{\text{KM}} = 209 - 1{,}15 \times \left(\frac{\#\text{Mots}}{\#\text{Phrases}}\right) - 68 \times \left(\frac{\#\text{Syllabes}}{\#\text{Mots}}\right)$$

où #Mots représente le nombre total de mots du tweet (y compris les hashtags), #Phrases désigne le nombre de phrases, et #Syllabes correspond au nombre total de syllabes. Le score résultant se situe généralement entre 0 et 100, les valeurs les plus élevées indiquant une meilleure facilité de lecture. Le décompte syllabique suit les conventions françaises standard, avec des ajustements spécifiques pour les terminaisons en « e » muet (ex. -e, -es) afin de refléter fidèlement le rythme oral et la réalité linguistique du contenu.

4.2 Flesch Reading Ease pour le contenu en anglais

Cet indice est particulièrement adapté aux textes courts dans la mesure où il ne dépend d'aucun dictionnaire de fréquences externe[11]. Il repose sur deux paramètres principaux : la longueur moyenne des phrases et la densité syllabique moyenne des mots. La formule est appliquée comme suit :

$$RE = 206{,}835 - 1{,}015 \left(\frac{\#\text{Mots}}{\#\text{Phrases}}\right) - 84{,}6 \left(\frac{\#\text{Syllabes}}{\#\text{Mots}}\right)$$

4.3 Évaluation basée sur les LLM (Llama-3.3-70B)

Bien que les métriques de lisibilité traditionnelles restent largement utilisées, elles demeurent limitées pour capturer la véritable difficulté linguistique, car elles reposent principalement sur des caractéristiques de surface telles que la longueur des mots et la structure des phrases. Pour pallier ces limites, les approches automatiques ont intégré des indicateurs linguistiques plus profonds, notamment la fréquence des mots, l'analyse morphologique et la structure syntaxique, ce qui a conduit au développement de modèles d'apprentissage automatique capables d'évaluer dynamiquement la lisibilité.
Plus récemment, l'émergence des grands modèles de langage (LLM) a introduit un nouveau paradigme dans l'évaluation de la lisibilité. Contrairement aux formules traditionnelles, les LLM adoptent une perspective holistique en capturant à la fois la cohérence sémantique et les relations discursives implicites. Plusieurs travaux ont exploré le potentiel de différents LLM open-source à cet effet, couvrant une grande variété d'architectures, de tailles de modèles et de données d'entraînement, et montrant que les modèles de grande taille offrent un meilleur alignement avec les jugements humains, notamment pour les contenus courts et informels. Dans cette étude, nous retenons le modèle *Llama-3.3-70B*. Ce modèle évalue la lisibilité de courts textes selon une échelle ordinale à cinq niveaux, dans laquelle une valeur plus élevée indique une meilleure lisibilité : Très difficile (1), Difficile (2), Assez difficile (3), Facile (4), Très facile (5).
Afin d'optimiser la qualité des évaluations produites par Llama-3.3-70B, nous avons expérimenté trois prompts distincts. Le prompt sélectionné ancre l'évaluation dans une perspective pédagogique, cohérente avec le contexte des écoles de commerce étudiées. La conversion des évaluations qualitatives en valeurs numériques permet une comparaison directe avec les métriques classiques et facilite l'intégration dans le cadre multicritère SAW.

TABLE 1 – Échelles de niveaux de lisibilité

Niveau	Métriques traditionnelles		LLM
	FRE (anglais)	KM (français)	Llama-3.3-70B
Très difficile	0–29	0–29	1
Difficile	30–49	30–49	2
Assez difficile	50–69	50–69	3
Facile	70–89	70–89	4
Très facile	90–100	90–100	5

5 Résultats et analyse

Cette section présente une analyse comparative des scores et niveaux de lisibilité des tweets organisationnels obtenus à l'aide de Llama-3.3-70B et de formules traditionnelles telles que Flesch Reading Ease (FRE) et Kandel-Moles (KM). Les observations mettent en évidence la pertinence des outils basés sur l'IA pour évaluer la lisibilité de contenus courts, informels et bilingues des tweets sur X.

Pour permettre une comparaison quantitative, les évaluations qualitatives produites par Llama-3.3-70B ont été converties en valeurs numériques selon l'échelle du Tableau 1, où une valeur plus élevée indique une meilleure lisibilité perçue, en cohérence avec les indices FRE et KM.

5.1 Description du jeu de données

Le corpus utilisé dans cette étude comprend 11 000 tweets collectés auprès des comptes officiels d'écoles de commerce françaises sur la plateforme X. Ces institutions utilisent activement X comme canal de communication stratégique pour diffuser des informations relatives à leurs programmes académiques, leurs événements institutionnels et leurs partenariats. La diversité des registres employés entre communication formelle, annonces d'événements et interactions avec les étudiants fait de ce corpus un terrain d'analyse représentatif des enjeux de qualité rédactionnelle en contexte organisationnel.

Un sous-ensemble de 50 tweets a été sélectionné à partir de ce corpus et annoté par un expert en communication numérique, occupant actuellement le poste de responsable de la communication au sein d'une école de commerce et disposant de plus de 20 ans d'expérience dans la gestion des réseaux sociaux. La taille de ce sous-ensemble a été volontairement limitée afin de permettre une analyse experte détaillée, cohérente et fondée sur une évaluation qualitative approfondie. L'expert a joué deux rôles distincts et indépendants dans cette étude. D'une part, il a défini les poids de la méthode SAW à travers un processus structuré d'élicitation, réalisé préalablement et indépendamment de la tâche d'annotation. D'autre part, il a annoté les 50 tweets en fournissant des jugements de lisibilité servant de référence de vérité terrain pour l'évaluation.

5.2 Analyse comparative des scores de lisibilité

Le Tableau 2 présente une comparaison des scores de lisibilité pour un échantillon de tweets, en confrontant les scores produits par les métriques traditionnelles (FRE et Kandel-Moles), les évaluations du modèle Llama-3.3-70B ainsi que les jugements de l'expert.

Dans l'ensemble, la classification produite par Llama-3.3-70B est cohérente avec celle de l'expert, notamment pour les tweets dont le niveau de difficulté est clairement défini. Par exemple, les tweets T_1, T_2, T_3 et T_9 sont jugés « Très facile » à la fois par Llama-3.3-70B et par l'expert. De même, les tweets T_7 et T_8 sont tous deux évalués comme « Facile » par le modèle et par l'expert. Le tweet T_5 est identifié comme « Difficile » par Llama-3.3-70B ainsi que par l'expert. Notons toutefois que le score FRE associé (8,87) le classerait dans la catégorie « Très difficile » selon les seuils présentés dans le Tableau 1. Cet écart illustre les limites de l'indice FRE pour les contenus courts et contextuels.

Des divergences subsistent néanmoins pour les tweets T_6, T_{10}, T_{28}, T_{43} et T_{50}. Le cas du Tweet 50 est particulièrement illustratif : son score KM (52,75) indique une difficulté intermédiaire selon la formule classique, tandis que l'expert l'évalue comme « Facile » et Llama-3.3-70B lui attribue un score de 3 (« Assez difficile »). Cette divergence entre les métriques automatiques d'une part, et le jugement de l'expert d'autre part, révèle que les formules traditionnelles peinent à rendre compte de la lisibilité perçue dans des contenus courts, ce qui met en évidence la nécessité de combiner l'automatisation et l'expertise humaine afin d'obtenir une évaluation plus complète et plus robuste de la qualité de présentation textuelle.

TABLE 2 – Comparaison des scores de lisibilité pour un échantillon de tweets

Tweet	FRE	KM	Llama-3.3-70B	Jugement expert
T_1	–	75,40	5	Très facile
T_2	–	72,15	5	Très facile
T_3	–	78,30	5	Très facile
T_4	–	45,20	3	Assez difficile
T_5	8,87	–	2	Difficile
T_6	10,13	–	1	Assez difficile
T_7	–	48,90	4	Facile
T_8	–	55,60	4	Facile
T_9	–	82,10	5	Très facile
T_{10}	12,16	–	1	Difficile
⋮	⋮	⋮	⋮	⋮
T_{28}	18,45	–	1	Assez difficile
T_{43}	–	41,30	3	Très difficile
⋮	⋮	⋮	⋮	⋮
T_{50}	–	52,75	3	Facile

5.3 Interprétation des résultats

La distribution des scores de lisibilité révèle des comportements distincts entre les trois approches. Les métriques traditionnelles reposent sur des caractéristiques linguistiques de surface telles que la longueur des phrases ou la densité

syllabique. En revanche, Llama-3.3-70B intègre des informations sémantiques et contextuelles, lui permettant d'évaluer la lisibilité de façon plus proche de la perception humaine. Cette distinction reflète la différence entre une mesure de la forme syntaxique du texte (approche classique) et une évaluation de son sens perçu dans un contexte donné (approche LLM).

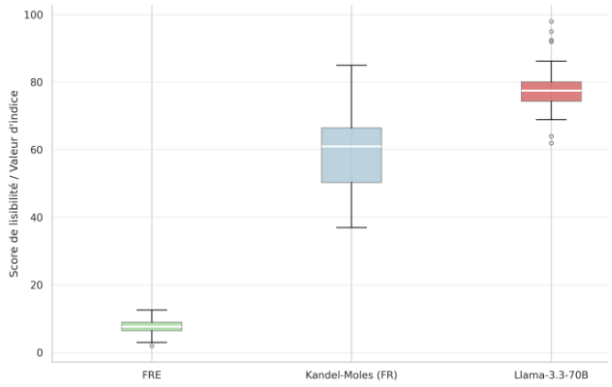


FIGURE 2 – Distribution des scores de lisibilité

5.4 Caractéristiques textuelles spécifiques à X

Les tweets contiennent des éléments textuels spécifiques qui les distinguent des textes écrits traditionnels, notamment les emojis, les hashtags, les mentions (@) et les hyperliens. Ces éléments sont largement utilisés pour transmettre des émotions, structurer l'information ou interagir avec d'autres utilisateurs. Dans le contexte de l'évaluation de la lisibilité, ces caractéristiques peuvent influencer la compréhension des tweets.

Emojis : Les emojis sont des symboles visuels utilisés pour enrichir les messages textuels en fournissant des indices contextuels ou émotionnels supplémentaires. Afin d'évaluer l'influence des emojis sur la lisibilité des tweets, ces derniers ont été regroupés en cinq catégories définies par l'expert : 0, 1, 2, 3 et $>$ 3 emojis. Le Tableau 3 présente, pour chaque catégorie, le nombre de tweets, le score moyen de lisibilité (SM), ainsi que les écarts-types (σ) correspondants.

TABLE 3 – Scores de lisibilité selon le nombre d'emojis

Nombre d'emojis	Nombre de tweets	SM	Écart-type
0 emoji	9 050	2,38	0,38
1 emoji	1 310	2,46	0,30
2 emojis	351	2,50	0,28
3 emojis	137	2,50	0,26
$>$ 3 emojis	109	2,42	0,46
Total	**11 000**	**2,40**	**0,37**

Les scores de lisibilité restent globalement stables pour toutes les catégories, avec une variation de 2,38 à 2,50 sur l'échelle 1–5. Les tweets contenant 2 ou 3 emojis présentent un score légèrement supérieur ($SM = 2{,}50$), tandis que les tweets sans emoji affichent le score le plus faible ($SM = 2{,}38$). Toutefois, ces écarts demeurent minimes et ne permettent pas de conclure à un effet significatif du nombre d'emojis sur la lisibilité perçue. Les faibles écarts-types (compris entre 0,26 et 0,46) témoignent d'une relative homogénéité des scores au sein de chaque catégorie, à l'exception de la catégorie $>$ 3 emojis**,** qui présente une dispersion légèrement plus élevée ($\sigma = 0{,}46$), imputable à la plus grande variabilité des tweets fortement chargés en emojis.

Hashtags : Les hashtags sont utilisés dans les tweets comme marqueurs de métadonnées permettant de référencer des sujets ou des thèmes et de rendre le contenu facilement découvrable sur les réseaux sociaux. Bien qu'ils permettent de catégoriser l'information et d'améliorer la visibilité du contenu, un nombre excessif de hashtags peut réduire la clarté du message et affecter négativement sa lisibilité. Dans le cadre de cette étude, les tweets ont été regroupés en fonction du nombre de hashtags qu'ils contiennent, selon une classification définie avec l'expert.

Le Tableau 4 présente, pour chaque catégorie, le nombre de tweets et l'écart de score de lisibilité par rapport aux tweets sans hashtag, calculé sur les trois méthodes.

TABLE 4 – Écarts de score de lisibilité selon les catégories de hashtags

Nombre de hashtags	Nombre de tweets	FRE	KM	Llama-3.3-70B
0	5 221	-	-	-
1	1 898	−0,47	−0,43	−0,12
2–4	2 808	−0,56	−0,51	−0,18
≥ 5	1 073	−0,76	−0,69	−0,27
Total	**11 000**			

La Figure 3 illustre la distribution des écarts de score de lisibilité pour les trois approches étudiées (FRE, Kandel-Moles et Llama-3.3-70B), calculée sur l'ensemble des **5 779** tweets contenant au moins un hashtag (**52,54** % du corpus). On observe que la majorité de la distribution se concentre dans les valeurs négatives pour les trois méthodes, ce qui suggère que l'inclusion de hashtags tend globalement à complexifier la structure du message et à réduire le score de lisibilité.

Toutefois, les écarts observés diffèrent selon la méthode : FRE et Kandel-Moles produisent des écarts plus importants que Llama-3.3-70B, qui présente une distribution plus stable. Cela s'explique par le fait que les métriques classiques pénalisent davantage les mots concaténés sans espaces, caractéristique des hashtags, tandis que Llama-3.3-70B est capable d'en inférer le sens dans le contexte global du tweet.

Bien que la présence de hashtags puisse modifier de manière significative le score de lisibilité d'un tweet, leur im-

pact sur la distribution globale de l'échantillon demeure relativement limité, notamment pour Llama-3.3-70B.
À titre d'illustration, considérons le tweet suivant :

> *#BusinessSchool #MBA #HigherEducation #Leadership — Join our MBA program and boost your career. Applications are now open !*

Pour ce type de tweet, l'inclusion des hashtags entraîne un écart moyen de $\Delta = -0{,}76$ point selon les métriques classiques, contre $\Delta = -0{,}27$ point pour Llama-3.3-70B, par rapport au texte brut. Cette comparaison illustre que les hashtags complexifient davantage la structure formelle mesurée par FRE et Kandel-Moles, tout en enrichissant la sémantique thématique perçue par Llama-3.3-70B.

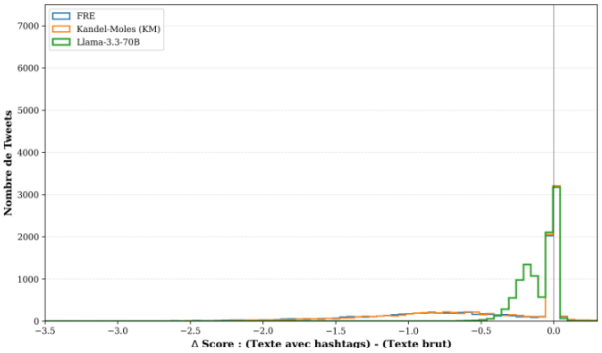


FIGURE 3 – Distribution des écarts de score de lisibilité selon le nombre de hashtags

6 Analyse de corrélation

Pour évaluer l'alignement entre les métriques automatisées et les préférences de l'expert, une matrice de corrélation complète a été calculée en utilisant à la fois le coefficient de Pearson (r) et celui de Spearman (ρ). Ces deux mesures permettent de quantifier le degré d'accord entre les évaluations automatisées et les jugements de l'expert, en couvrant à la fois les relations linéaires et les relations monotones ordinales.
Le coefficient de corrélation produit-moment de Pearson (r) mesure la relation linéaire entre deux variables continues :

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{8}$$

où x_i et y_i représentent les valeurs individuelles de chaque méthode, $\bar{x}$ et $\bar{y}$ leurs moyennes respectives, et N le nombre total d'observations.
Compte tenu de la nature ordinale des niveaux de lisibilité, le coefficient de corrélation de rang de Spearman (ρ) a également été employé :

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{9}$$

où d_i est la différence entre les rangs de chaque observation i, et N le nombre total d'observations.

La complémentarité de ces deux coefficients garantit une évaluation robuste : le coefficient de Pearson capture la cohérence des écarts numériques, tandis que le coefficient de Spearman valide la stabilité de l'ordre de classement, dimension cruciale pour les systèmes d'aide à la décision ordinaux.

TABLE 5 – Corrélations entre les jugements de l'expert et les caractéristiques textuelles

Caractéristique / Modèle	**Pearson** r	**Spearman** ρ
Flesch Reading Ease	0,17	0,22
Kandel-Moles	0,21	0,23
Nombre de hashtags	−0,23	−0,05
Nombre d'emojis	0,04	0,09
Efficacité des emojis	0,11	0,13
Mots par phrase	−0,16	−0,27
Hapax legomena (%)	0,36	0,36
Fautes d'orthographe (%)	0,27	0,21
Parenthèses (%)	−0,30	−0,27
Abréviations (%)	−0,03	−0,11
Références croisées (%)	−0,15	−0,21
Difficulté grammaticale	−0,09	−0,14
Polysémie (%)	−0,04	−0,13
Llama-3.3-70B	**0,52**	**0,50**

Les résultats présentés dans le Tableau 5 confirment que la lisibilité des tweets organisationnels constitue un construit multidimensionnel. Le modèle *Llama-3.3-70B* se distingue nettement ($r = 0{,}52$, $\rho = 0{,}50$), en atteignant un niveau d'alignement avec les jugements de l'expert supérieur à celui obtenu par les métriques classiques. La faible corrélation de FRE ($r = 0{,}17$) souligne son inadéquation structurelle pour les contenus courts, où la longueur des phrases est contrainte par la limite de caractères de la plateforme. La corrélation modérée de KM ($r = 0{,}21$) reflète sa meilleure adaptation aux spécificités phonétiques du français, mais confirme néanmoins ses limites pour les contenus informels et multimodaux des réseaux sociaux.
Parmi les indicateurs textuels complémentaires, les *hapax legomena* ($r = 0{,}36$) constituent le prédicteur de complexité le plus fort après *Llama-3.3-70B*, confirmant que la diversité lexicale est un signal pertinent de la difficulté perçue. À l'inverse, la présence de parenthèses ($r = -0{,}30$) est négativement corrélée avec les jugements de l'expert, confirmant que ces éléments perturbent la fluidité de lecture dans le format court du tweet.
Par ailleurs, l'analyse des caractéristiques spécifiques à la plateforme X révèle des résultats nuancés. Le nombre de hashtags ($r = -0{,}23$) est négativement corrélé avec les jugements de l'expert, ce qui confirme que leur usage excessif nuit à la lisibilité perçue dans le format court du tweet. En revanche, les emojis présentent un effet inverse. Leur présence montre une corrélation légèrement positive ($r = 0{,}04$, $\rho = 0{,}09$) avec les jugements de l'expert, suggé-

rant qu'un usage modéré des emojis contribue positivement à la lisibilité perçue.

7 Évaluation globale de la qualité de présentation textuelle par SAW

Cette section présente l'application numérique de la méthode SAW à un échantillon représentatif de tweets issus du corpus. Le calcul se déroule en quatre étapes : construction de la matrice de décision, normalisation, agrégation pondérée et classement final.

7.1 Définition des critères et des poids

Les trois critères retenus et leurs poids W_j ont été définis lors de la session de validation avec l'expert, conformément à la contrainte de normalisation $\sum_{j=1}^{3} W_j = 1$.
Le Tableau 6 présente ces critères et les poids associés.

TABLE 6 – Critères d'évaluation et poids définis par l'expert (HITL)

	Critère	**Type**	**Poids** W_j
g_1	Lisibilité	Bénéfice	0,60
g_2	Présence des hashtags	Bénéfice	0,20
g_3	Usage des emojis	Bénéfice	0,20

7.2 Normalisation de la matrice de décision

Une normalisation des scores bruts est nécessaire afin de les rendre comparables sur une échelle commune $[0, 1]$. Chaque performance brute x_{ij} est normalisée selon la nature du critère g_j, conformément à l'Équation (4). Dans le cas présent, tous les critères retenus étant de type bénéfice, une valeur plus élevée indiquant une meilleure qualité, la normalisation s'effectue par rapport au maximum observé sur chaque critère.
À titre d'illustration, les deux cas extrêmes du critère g_1 sont présentés ci-dessous, couvrant respectivement le tweet le plus lisible et le moins lisible de l'échantillon. Le maximum observé sur ce critère est $x_1^* = \max_i\{x_{i1}\} = 5{,}0$ (score Llama-3.3-70B maximal, niveau Très facile).
Pour le tweet T_9, qui présente la lisibilité maximale ($x_{9,1} = 5{,}0$, niveau Très facile) :

$$\tilde{x}_{9,1} = \frac{x_{9,1}}{x_1^*} = \frac{5{,}0}{5{,}0} = 1{,}000 \tag{10}$$

Pour le tweet T_6, dont le score Llama-3.3-70B est le plus faible de l'échantillon ($x_{6,1} = 1{,}0$, niveau Très difficile) :

$$\tilde{x}_{6,1} = \frac{x_{6,1}}{x_1^*} = \frac{1{,}0}{5{,}0} = 0{,}200 \tag{11}$$

Ce calcul est appliqué de manière identique à l'ensemble des tweets et pour chacun des trois critères g_1, g_2 et g_3. La matrice normalisée complète $\tilde{X} = (\tilde{x}_{ij}) \in [0, 1]^{m\times 3}$, avec m tweets en lignes et $n = 3$ critères en colonnes, est présentée dans le Tableau 7.

7.3 Agrégation pondérée et classement

Le score global S_i de chaque tweet T_i est calculé conformément à l'Équation (7). Le Tableau 7 présente les scores normalisés $\tilde{x}_{ij}$ et le score global S_i pour l'échantillon de tweets évalués.

TABLE 7 – Scores SAW appliqués à l'échantillon de tweets

Tweet	$\tilde{x}_{i1}$	$\tilde{x}_{i2}$	$\tilde{x}_{i3}$	S_i
T_1	1,000	0,600	1,000	0,920
T_2	1,000	1,000	0,556	0,911
T_3	1,000	1,000	0,556	0,911
T_4	0,600	1,000	0,556	0,671
T_5	0,400	1,000	0,778	0,596
T_6	0,200	1,000	0,556	0,431
T_7	0,800	1,000	0,556	0,791
T_8	0,800	1,000	0,556	0,791
T_9	1,000	1,000	1,000	**1,000**
T_{10}	0,200	1,000	0,556	0,431
⋮	⋮	⋮	⋮	⋮
T_{28}	0,200	1,000	0,556	0,431
T_{43}	0,600	1,000	0,556	0,671
⋮	⋮	⋮	⋮	⋮
T_{50}	0,600	1,000	0,556	0,671

Le tweet T_9 obtient le score maximal ($S_9 = 1{,}000$), reflétant une excellence sur les trois critères : lisibilité maximale évaluée par Llama-3.3-70B ($\tilde{x}_{9,1} = 1{,}000$, niveau Très facile), le score des hashtags optimale ($\tilde{x}_{9,2} = 1{,}000$) et le score des emojis maximale ($\tilde{x}_{9,3} = 1{,}000$), en cohérence avec le jugement expert *Très facile*.
À l'inverse, les tweets T_6, T_{10} et T_{28} présentent les scores les plus faibles ($S_i = 0{,}431$), en raison d'une lisibilité très limitée ($\tilde{x}_{i1} = 0{,}200$, score Llama-3.3-70B = 1, niveau Très difficile).
Ces résultats confirment l'effet de compensation partielle inhérent à la méthode SAW : une faiblesse sur le critère dominant g_1 (lisibilité, $W_1 = 0{,}60$) ne peut être que partiellement compensée par les critères secondaires g_2 et g_3. La décomposition additive du score S_i offre ainsi une explicabilité directement opérationnelle, permettant à l'expert de formuler des recommandations ciblées en vue d'optimiser la communication institutionnelle sur la plateforme X.

8 Conclusion

Dans cet article, nous avons présenté une approche hybride centrée sur l'humain pour l'évaluation explicable de la qualité de présentation textuelle des tweets organisationnels. En combinant l'expertise humaine avec les capacités d'analyse du modèle Llama-3.3-70B au sein d'un cadre multicritère SAW, nous avons formalisé trois critères pertinents lisibilité, la présence des hashtags et usage des emojis et

produit des scores de qualité à la fois robustes, explicables et contextualisés.

Les résultats expérimentaux montrent que le modèle Llama-3.3-70B surpasse les métriques classiques en capturant efficacement les nuances sémantiques et contextuelles des messages courts, avec une corrélation de Pearson $r = 0{,}52$ avec les jugements de l'expert. Cette capacité à aligner les scores automatiques sur l'évaluation humaine témoigne de la pertinence d'un cadre hybride qui tire parti à la fois de la cohérence des modèles de langue et du jugement expert. Par ailleurs, l'analyse des éléments spécifiques à la plateforme X (hashtags et emojis) a confirmé leur rôle non négligeable sur la lisibilité et la complexité perçue des messages : leur prise en compte explicite dans le modèle multicritère améliore la finesse de l'évaluation par rapport à des approches purement lexicales. Cette approche offre ainsi un compromis entre automatisation et compréhension fine du contenu, répondant directement aux besoins des décideurs pour la prise de décision stratégique en matière de communication numérique.

Enfin, cette recherche ouvre la voie à plusieurs perspectives. Tout d'abord, l'intégration de la qualité multimodale (image, vidéo) et la validation sur d'autres types d'organisations (collectivités, ONG, institutions publiques) élargiront la portée et la généralité de l'approche. Ensuite, l'extension vers l'évaluation en temps réel des flux de données sur les réseaux sociaux ouvrirait la voie à des systèmes de veille et d'aide à la décision opérationnels. En combinant le raisonnement humain et l'intelligence artificielle, cette approche constitue une avancée significative vers des systèmes de support décisionnel plus explicables et fiables dans l'analyse des données des réseaux sociaux à des fins organisationnelles.

Références

[1] I. Fotovat Ahmadi, A. Waltenrath, and C. Janze. Congruency and users' sharing on social media platforms : A novel approach for analyzing content. *Journal of Advertising*, 51(4) :489–507, 2022.

[2] A. Albladi, M. Islam, and C. Seals. Sentiment analysis of twitter data using nlp models : A comprehensive review. *IEEE Access*, 8 :122199–122219, 2020.

[3] B. Basaran. Enhancing readability assessment for language learners : A comparative study of ai and traditional metrics in german textbooks. *European Journal of Educational Research*, 15(1) :101–119, 2025.

[4] I. Cachola, D. Khashabi, and M. Dredze. Evaluating the evaluators : Are readability metrics good measures of readability ? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[5] E. Dale and J. S. Chall. The concept of readability. *Elementary English*, 26(1) :19–26, 1949.

[6] S. W. Davis, C. Horváth, A. Gretry, and N. Belei. Say what ? how the interplay of tweet readability and brand hedonism affects consumer engagement. *Journal of Business Research*, 100 :150–164, 2019.

[7] A. Gaydhani, V. Doma, S. Kendre, et al. Detecting hate speech and offensive language on twitter using machine learning : An n-gram and tf-idf based approach. *arXiv preprint arXiv :1809.08651*, 2018.

[8] D. C. Gkikas, K. Tzafilkou, P. K. Theodoridis, A. Garmpis, and M. C. Gkikas. How do text characteristics impact user engagement in social media posts : modeling content readability, length, and number of hashtags in facebook. *International Journal of Information Management Data Insights*, 2(1) :100067, 2022.

[9] S. Greco, B. Matarazzo, and R. Slowinski. Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research*, 129(1) :1–47, 2001.

[10] A. Ismail, K. S. Kuppusamy, A. Kumar, and P. K. Ojha. Connect the dots : Accessibility, readability and site ranking – an investigation with reference to top ranked websites of government of india. *Journal of King Saud University - Computer and Information Sciences*, 31(4) :528–540, 2019.

[11] P. Jacob and A. L. Uitdenbogerd. Readability of twitter tweets for second language learners. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 19–27, 2019.

[12] I. Kaliszewski and D. Podkopaev. Simple additive weighting – a metamodel for multiple criteria decision analysis methods. *Expert Systems with Applications*, 2016.

[13] L. Kandel and A. Moles. Application de l'indice de flesch à la langue française. *Cahiers d'Études de Radio-Télévision*, 19 :253–274, 1958.

[14] F. Khan, Z. Hafeez, A. Ijaz, and R. Shaheen. Predicting factors to get maximum social media engagement of customers for brand building. In *Proceedings of the Al Yamamah University Engineering Forum (YUENG)*, Riyadh, Saudi Arabia, 2019.

[15] M. L. Khan, M. Wasim, and A. Kaur. The role of emojis and hashtags in social media engagement : A sentiment and communication perspective. *Social Network Analysis and Mining*, 14(1) :1–15, 2024.

[16] J. P. Kincaid, R. P. Fishburne Jr., and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command, Research Branch, Millington, TN, USA, 1975.

[17] G. S. Kumar, N. Y. Reddy, R. G. Mudhiraj, and S. Ch. Prasad. Deep learning-based social media trend analyzer to predict trends over time. *Journal of Science Engineering Technology and Management Science*, 2(7S) :69–79, 2025.

[18] R. Kumar, R. Sinha, S. Saha, and A. Jatowt. Extracting the full story : A multimodal approach and dataset to crisis summarization in tweets. *IEEE Transactions on Computational Social Systems*, 2024.

[19] F. Marulli, L. Campanile, M. S. De Biase, S. Marrone, L. Verde, and M. Bifulco. Understanding readability of large language models output : An empirical analysis. *arXiv preprint*, 2024.

[20] W. Pan, X. Li, X. Chen, and R. Xu. Textual form features for text readability assessment. *Natural Language Processing*, 31(3) :800–841, 2025.

[21] G. M. Pascoal, M. van den Bosch, O. Viberg, and J. Wong. Improving text readability to support student comprehension and learning : An llm-powered approach. In *Two Decades of TEL : From Lessons Learnt to Challenges Ahead*, pages 291–305, September 2025.

[22] A. T. Rimadewi, Y. Azis, D. Sari, and I. Soemaryani. Social media reporting : How to do it right for strategic decision making. *Journalism and Media*, 6(4), 2025.

[23] I. Saad. Explainable AI and Multicriteria Decision Support : Real World Applications and Perspectives. In *Proceedings of EURO 2025*, pages 22–25, Leeds, United Kingdom, June 2025.

[24] C. Salvatore, S. Biffignandi, and A. Bianchi. Social media and twitter data quality for new social indicators. *Social Indicators Research*, 156 :601–630, 2021.

[25] I. Temnikova, S. Vieweg, and C. Castillo. The case for readability of crisis communications in social media. In *Proceedings of the 24th International Conference on World Wide Web Companion (WWW '15 Companion)*, pages 1245–1250, 2015.

[26] L. Zhang, J. Abhani, J. B., A. Yadav, M. S. Ab Yajid, F. Mowafaq, and S. Vats. Automatic text readability assessment for educational content based on graph representation learning. *Scientific Reports*, 2026.

Effets d'ancrage liés à la pression temporelle et à la difficulté des tâches dans un contexte de décision humaine assistée par IA

Jules Leguy[1], Nicolas Soulié[2], Felipe Torres Figueroa[2], Andon Tchechmedjiev[1], Jacky Montmain[1], Sébastien Harispe[1]

[1] SyCoIA, IMT Mines Ales, Ales, France
[2] Université Paris-Saclay, Univ Evry, IMT-BS, LITEM, 91025, Evry-Courcouronnes, France

jules.leguy@mines-ales.fr

Résumé

Les systèmes de décision assistée par IA se multiplient dans les domaines à forts enjeux. Une calibration adéquate de la confiance et de la reliance (inclination à suivre les prédictions de l'IA) des opérateurs humains envers ces systèmes est nécessaire pour en permettre un usage approprié. Cette calibration peut être affectée par la pression temporelle et la difficulté des tâches, ainsi que par d'éventuels effets d'ancrage engendrés par les premières interactions avec le système. Nous menons une étude permettant d'examiner l'impact des conditions d'exposition initiale à un système de décision assistée par IA sur le comportement ultérieur des utilisateurs (N = 172). Nous montrons que les conditions de pression temporelle et de difficulté de la première tâche peut influencer la sur-reliance, la confiance, la charge de travail et les temps de décision sur l'ensemble de l'interaction avec le système prédictif. Une analyse plus fine montre que le facteur déterminant est la première rencontre avec une condition exigeante, qui déclenche un changement de comportement persistant.

Mots-clés

IA centrée sur l'humain, Décision assistée par IA, IHM

Abstract

AI-assisted decision systems are proliferating in high-stakes domains. Proper calibration of human operators' trust and reliance (inclination to follow AI predictions) on these systems is necessary to ensure appropriate use. This calibration may be affected by time pressure and task difficulty, as well as by potential anchoring effects arising from initial interactions with the system. We conduct a study examining the impact of initial exposure conditions to an AI-assisted decision system on subsequent user behavior (N = 172). We show that the time pressure and difficulty conditions of the first task can influence over-reliance, trust, workload, and decision times throughout the interaction with the predictive system. Further analysis shows that the determining factor is the first encounter with a demanding condition, which triggers a persistent behavioral shift.

Keywords

Human-centered AI, AI-assisted decision making, HCI

1 Introduction

Les systèmes d'aide à la décision basés sur l'intelligence artificielle (IA) connaissent une croissance rapide dans de nombreux domaines à forts enjeux. En médecine, des modèles prédictifs assistent des radiologues ou des dermatologistes dans la détection de lésions cancéreuses [10, 7]. Dans le domaine militaire, ces systèmes sont déployés pour la fusion de données de renseignement et la planification opérationnelle [20].

Un système d'aide à la décision par IA est un système qui fournit à un opérateur humain une recommandation ou une prédiction issue d'un modèle algorithmique, afin de l'assister dans l'élaboration de son jugement pour traiter des instances d'un problème donné. Dans ce paradigme, la responsabilité de la décision incombe toujours à l'humain. Il est donc essentiel que la confiance et la reliance de l'opérateur envers le système soient calibrées en accord avec les performances effectives de ce dernier [18]. Nous faisons la distinction entre la confiance qui est une attitude envers le système prédictif, et la reliance qui est le comportement visant à reproduire les mêmes décisions que celles qui sont suggérées par le système prédictif, ou des décisions similaires [25]. Une confiance trop faible conduit à rejeter des recommandations pertinentes, tandis qu'une confiance excessive pousse l'opérateur à adopter les recommandations du système sans jugement critique, même lorsqu'elles sont erronées [18]. Dans cet article, nous utilisons le terme de sur-reliance pour désigner ce dernier comportement. Par ailleurs, le contexte opérationnel dans lequel s'inscrivent ces interactions peut influencer significativement la qualité des décisions. La pression temporelle et la difficulté des tâches qui sont traitées par les opérateurs sont susceptibles d'avoir un impact direct sur l'adoption des recommandations du système [23, 31]. A titre d'exemple, les radiologues travaillent souvent dans un contexte de stress intense, où ils doivent analyser avec une extrême précision des images médicales aux enjeux potentiellement vitaux, sous une forte pression temporelle et dans un environnement fréquemment

perturbé par les urgences, les interruptions et la charge de travail élevée [3]. En contexte militaire, la pression temporelle et la complexité des tâches sont des caractéristiques intrinsèques des environnements opérationnels, où les décideurs doivent traiter des informations incomplètes dans des délais souvent très contraints [8], avec des répercussions également susceptibles d'être vitales.

Au-delà de ces facteurs contextuels, des mécanismes cognitifs comme l'effet d'ancrage – tendance à se fier excessivement à la première information reçue (appelée *ancre*) pour prendre une décision ou faire une estimation, même si cette information est arbitraire ou peu pertinente – peuvent également avoir un impact important sur la calibration de la confiance et de la reliance dans les systèmes prédictifs [12]. L'effet d'ancrage se manifeste dans les interactions humain-IA à deux niveaux distincts. Le premier niveau est celui des décisions individuelles, où la prédiction du modèle peut constituer une ancre cognitive dont il est difficile pour l'opérateur de s'affranchir, même lorsqu'il dispose d'indices contradictoires [22, 4]. La littérature propose à cet égard des *cognitive forcing functions* (interventions délibérément conçues pour interrompre le raisonnement automatique et promouvoir un traitement analytique) comme mécanismes permettant de limiter cet ancrage [4]. Le second niveau d'apparition des effets d'ancrage est celui des utilisations répétées d'un système prédictif. Un effet d'ancrage peut se former à partir de la première impression que l'opérateur se forge des performances globales du modèle, et persister lors de tâches ultérieures : des utilisateurs exposés à de bonnes performances initiales auront une plus forte tendance à la sur-reliance lors des interactions suivantes, tandis que ceux confrontés à des erreurs précoces auront tendance à sous-estimer les compétences du système [21, 28, 2, 11]. Swaroop et al. ont observé un lien entre pression temporelle et effet d'ancrage, selon lequel l'exposition à une condition de forte pression temporelle crée un ancrage qui augmente la reliance et rend les décisions des utilisateurs plus rapides [27]. Dans cet article, nous cherchons à confirmer et à étendre ces résultats en examinant conjointement les effets de la pression temporelle et de la difficulté des tâches.

Notre question de recherche est : quels sont les impacts de l'ordre d'exposition à différents niveaux de pression temporelle et de difficulté de tâche, sur le comportement d'opérateurs humains envers un système d'aide à la décision par IA ? Les données présentées dans cet article ont été obtenues en réalisant le protocole pré-enregistré à l'adresse `https://osf.io/56wnj/`, dont l'objet d'étude initial porte sur le comportement humain en présence de différentes techniques d'explicabilité de l'IA (XAI). Nous proposons une analyse post-hoc d'un sous-ensemble des données collectées, qui permet d'aborder la question de recherche considérée dans cet article. Nous reportons ainsi les résultats d'une étude expérimentale ($N = 172$) dans laquelle nous faisons varier le niveau de pression temporelle (faible ou forte) et de difficulté (faible ou élevée) pour la résolution d'une tâche de détection de motifs. L'étude utilise un plan mixte inter-intra sujets : une partie des participants résout les tâches de manière autonome, l'autre est assistée par un modèle prédictif. Tous les participants sont soumis dans un ordre aléatoire aux 4 couples de conditions combinant pression temporelle et difficulté. En isolant les effets des premières conditions expérimentales sur le comportement global des participants, nous étudions en particulier la reliance et la sur-reliance envers le modèle, la confiance déclarée, la charge de travail perçue et les temps de décisions des participants.

2 Contexte scientifique

2.1 Calibration de la confiance et reliance

L'un des enjeux majeurs de la collaboration humain-IA réside dans la calibration de la confiance et de la reliance de l'opérateur envers le système prédictif [30]. Il est important de distinguer la confiance envers le système prédictif, qui est une attitude ou une intention humaine et qui est mesurée de façon déclarative, de la reliance (comportement de s'appuyer sur les prédictions du modèle pour la prise de décision) qui est un comportement effectif mesuré par des mesures d'accord avec les décisions du modèle [18, 25]. Bien que ces deux concepts soient a priori liés, les humains peuvent avoir un comportement de reliance effective contradictoire avec la confiance qu'ils déclarent subjectivement [29]. Cette discordance peut être causée par des facteurs externes, telles que la volonté de gagner du temps ou de réduire la charge de travail [17]. Schaffer et al. observent que des sujets affichant une familiarité élevée avec la tâche réalisée déclarent également une confiance supérieure dans le système prédictif, tout en rejetant plus souvent les conseils de l'IA [24]. La calibration de la confiance et de la reliance est un processus complexe, pouvant être influencé par de nombreux facteurs. Parmi ceux-ci, les conditions de pression temporelle, la difficulté des tâches et les effets d'ancrage jouent un rôle important. Ces différents aspects sont examinés dans la suite de cette section.

2.2 Effets de la pression temporelle et de la difficulté des tâches

Lorsque les opérateurs humains doivent prendre une décision sous une pression temporelle forte, ils sont plus enclins à adopter les suggestions du système prédictif [26]. Cette tendance s'observe y compris lorsque les suggestions sont incorrectes, menant à un phénomène de sur-reliance [23]. De plus, cette pression peut engendrer une anxiété qui altère les capacités de résolution de la tâche, provoquant un phénomène de « suffocation » (choking) où l'utilisateur se précipite pour terminer le travail en déléguant aveuglément la décision à l'IA, alors même que le temps imparti aurait été suffisant en l'absence de stress [13]. Le second paramètre à considérer est la complexité de la tâche à réaliser. Lorsque les opérateurs humains considèrent une tâche comme difficile, ils sont également plus enclins à adopter les suggestions de l'IA [1]. Il a été démontré expérimentalement que la sur-reliance augmente de manière significative avec la difficulté de la décision [31]. La pression temporelle et la difficulté des tâches ont des effets similaires sur l'aug-

mentation de la sur-reliance, mais présentent des nuances théoriques dans leurs mécanismes. La difficulté correspond au niveau de ressources cognitives requis pour résoudre la tâche, alors que la pression temporelle réduit les capacités cognitives mobilisables en induisant un stress environnemental [5]. Peu de travaux ont étudié l'interaction entre la pression temporelle et la difficulté des tâches. Hermanns et Teubner observent que dans un contexte de pression temporelle forte, l'effet négatif d'une augmentation de la difficulté sur la performance est faible [16]. Ils expliquent ce résultat par le fait qu'en situation de forte pression temporelle, les utilisateurs ont déjà tendance à déléguer leur jugement aux prédictions du modèle, ce qui atténue l'effet additionnel du niveau de difficulté de la tâche.

2.3 Effets d'ancrage

L'effet d'ancrage est un biais cognitif fort qui se manifeste lorsqu'une valeur initiale, appelée ancre, exerce une influence disproportionnée sur des jugements ultérieurs [12]. Dans un contexte de décision assistée par IA, des effets d'ancrage peuvent se manifester à plusieurs niveaux. Ils peuvent apparaître au niveau des décisions individuelles, notamment lorsque les suggestions des modèles de décision sont soumises à l'opérateur humain en même temps que l'instance du problème à résoudre. La prédiction agit alors comme une ancre dont l'humain peut éprouver des difficultés à se défaire [22], d'autant plus que l'IA peut être perçue comme une figure d'autorité ou une source de données hautement objectives [6]. Des stratégies ont été proposées dans la littérature pour limiter l'effet d'ancrage lié aux prédictions individuelles, en augmentant délibérément le temps alloué à la décision [22] ou en intégrant des mécanismes dont le but est de forcer la réflexion humaine (*cognitive forcing functions*) avant de prendre connaissance de la prédiction du modèle [4]. Des effets d'ancrage peuvent aussi se manifester à une échelle temporelle plus large, entre plusieurs phases d'interactions avec le système prédictif. Une première impression positive du système, résultant d'une exposition initiale à des succès de l'IA, peut augmenter la reliance lors d'interactions ultérieures [28]. Biswas et al. observent un effet d'« inertie des croyances », par lequel une confiance élevée dans le système prédictif peut se transférer sous forme de confiance a priori pour une tâche ultérieure, y compris si les domaines des tâches sont différents [2]. La distribution temporelle des erreurs constitue également un facteur d'ancrage. Freel et al. observent que des erreurs précoces peuvent être plus facilement pardonnées par les utilisateurs que des erreurs tardives, car un mécanisme de réparation de la confiance peut avoir lieu si le système prédictif se comporte de manière fiable par la suite [11].

2.4 Positionnement de l'étude

Dans cet article, nous cherchons à mesurer les effets d'ancrage liés spécifiquement aux conditions de pression temporelle et de difficulté des tâches lors des premières interactions avec le modèle. Nous cherchons à étudier l'impact de ces effets sur la calibration de la confiance et de la reliance envers les modèles prédictifs. Les travaux de la littérature les plus proches sont ceux de Swaroop et al. [27], qui étudient le compromis entre efficacité (minimiser les temps de réponse) et performance (taux de bonnes réponses) pour la résolution d'une tâche dans un contexte de décision assistée par IA. Leur protocole fait varier la pression temporelle imposée aux participants. Ils montrent que la pression temporelle augmente la sur-reliance envers le système. Ils rapportent également (sans en présenter les données) que la première exposition à une tâche sous forte pression temporelle peut induire un comportement de sur-reliance persistant, qui se maintient lorsque la pression temporelle diminue par la suite. Notre protocole est similaire en ce que les participants sont exposés successivement à différents niveaux de pression temporelle, mais il s'en distingue sur plusieurs points. D'une part, nous étudions également l'effet de la difficulté des tâches [1] ; et d'autre part, nous analysons, en plus de la performance, du temps de décision et de la sur-reliance, qui sont les variables reportées par Swaroop et al., les variables subjectives que représentent la confiance déclarée envers le modèle et la charge de travail associée aux tâches. Enfin, notre analyse isole explicitement l'effet de la première condition expérimentale sur l'ensemble du comportement ultérieur des participants.

3 Méthodes

Cette section présente la méthodologie mise en œuvre pour étudier les effets d'ancrage liés à la pression temporelle et à la difficulté des tâches dans un contexte de décision assistée par IA. Comme indiqué dans l'introduction, les résultats présentés proviennent d'une analyse post-hoc de données collectées dans le cadre d'un protocole plus large incluant des techniques d'explicabilité (XAI), qui fera l'objet d'une communication dédiée. Nous restreignons notre analyse aux participants n'ayant pas eu accès à des explications, répartis en deux conditions : résolution sans assistance (groupe H) et avec une prédiction de l'IA (groupe H+IA).

3.1 Plan d'expériences

Le protocole expérimental a été soumis au Comité d'éthique de la recherche de l'Université de Montpellier (projet « Étude de l'impact de l'intelligence artificielle explicable sur la performance d'opérateurs humains pour une tâche de décision assistée par l'Intelligence Artificielle »), qui a rendu un avis favorable (Avis consultatif n°UM 2025-091). Il a aussi été pré-enregistré publiquement sur la plateforme Open Science Framework (OSF) [2] avant la collecte des résultats.

3.1.1 Tâche de décision

La tâche soumise aux participants consiste en une décision binaire : pour chaque couple (image, motif), le participant doit indiquer si le motif — ou l'une de ses rota-

1. Notons que Swaroop et al. considèrent également la difficulté comme une variable indépendante, mais uniquement pour la deuxième expérience reportée dans leur article. Seule leur première expérience traite de la question de l'ordre des conditions de pression temporelle.

2. Le protocole est accessible à l'adresse https://osf.io/56wnj/.

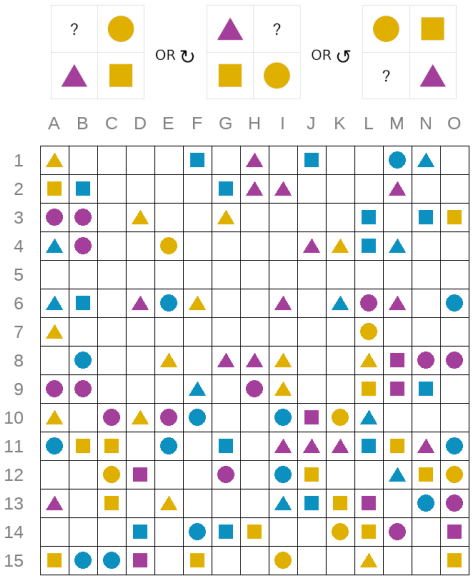


FIGURE 1 – Exemple de question soumise aux participants. Partie supérieure : motif ou ses rotations à identifier dans la grille. Partie inférieure : grille de symboles (de difficulté élevée). La question soumise explicitement est : le motif ou l'une de ses rotations est-il présent dans l'image ? Pour cet exemple, le motif est bien présent (la rotation vers la droite aux positions N11-N12-O12).

tions à 90° (gauche ou droite) — est présent dans l'image. Chaque image, générée aléatoirement, se présente sous la forme d'une grille de cellules contenant chacune un symbole coloré parmi un ensemble S^- ($|S^-| = 9$) ou aucun symbole ($\emptyset$), de sorte qu'une grille de taille $n \times n$ est un élément de $S^{n \times n}$ avec $S = S^- \cup \emptyset$. Un motif correspond à un élément de $S^{2\times2}$, et chaque image contient au plus une occurrence du motif ou de ses rotations. Une tâche comprend une série de questions portant sur différentes images et un même motif. La Figure 1 illustre une instance de question soumise aux participants.

Différentes conditions expérimentales sont considérées en fonction de la difficulté des tâches et de la pression temporelle imposée aux participants pour les résoudre. La difficulté est caractérisée comme fonction de la taille de la grille associée à l'image présentée à l'opérateur :

1. Facile : grille dans $S^{9\times9}$, motif dans $S^{2\times2}$.
2. Difficile : grille dans $S^{15\times15}$, motif dans $S^{2\times2}$.

La pression temporelle est caractérisée par le temps alloué aux participants pour prendre chaque décision. Un minuteur permet aux participants de savoir à tout moment combien de temps il leur reste pour répondre à une question. Lorsque le temps est imparti, l'interface passe à la question suivante. L'absence de réponse dans le temps imparti est considérée comme une mauvaise réponse. Deux niveaux de pression temporelle sont distingués :

1. Faible : 25 secondes maximum par image.
2. Forte : 10 secondes maximum par image.

Les différentes conditions expérimentales sont ainsi croisées de manière à proposer des tâches faciles ou difficiles avec une pression temporelle faible ou forte, au sens des définitions proposées ci-dessus. Nous comparons les performances et les comportement observés des participants ayant accès (groupe H+IA) ou non (groupe H) à l'assistance d'un modèle prédictif pour résoudre les tâches. L'assistance se présente sous la forme d'un encart "L'IA prédit : Oui/Non" visible dans l'interface de résolution (en anglais). Les participants ont dans tous les cas la responsabilité de la décision finale et sont libres de considérer ou d'ignorer les recommandations du modèle dans leurs processus décisionnels. Les participants ne sont pas informés de la performance des modèles, mais ils sont informés à la fin de chaque tâche de décision de leur score (pourcentage de réponses correctes). Ce choix vise à reproduire des conditions proches d'un usage réel, dans lequel les opérateurs construisent progressivement leur estimation des performances du système à partir de leur expérience accumulée lors d'utilisations répétées, et sans retour immédiat à chaque décision.

3.1.2 Variables dépendantes

Les variables dépendantes mesurées pour chaque tâche de l'expérience principale sont :

- Le score (proportion de réponses correctes, i.e. *accuracy*).
- La reliance (proportion de réponses identiques à celles de l'IA).
- La sur-reliance (proportion de réponses identiques à celles de l'IA lorsque celle-ci produit une mauvaise réponse).
- Le temps de réponse moyen.
- La confiance dans le système d'assistance (échelle de Likert à 7 points pour la question « Je fais confiance à l'IA pour détecter le motif et ses rotations. »).
- Charge de travail : somme de l'enquête standard NASA-TLX [14] sur une échelle à 7 points par question. La question concernant la charge physique est exclue de la somme car elle n'est pas considérée pertinente ici. Cette question a préalablement été exclue dans d'autres travaux de la littérature [28].

Les questionnaires pour mesurer la confiance et la charge de travail sont soumis aux participants immédiatement après les tâches de l'expérience principale, et avant de leur communiquer leur score pour la tâche courante, afin de ne pas biaiser la mesure de la charge de travail — dont l'une des questions porte sur l'estimation de la performance individuelle lors de la tâche.

3.1.3 Procédure

La procédure suivie par chaque participant comprend deux phases principales, illustrées en Figure 2. Une phase introductive permet d'abord aux participants de se familiariser avec l'interface et les tâches, afin de se préparer pour la phase d'expérience principale, pendant laquelle les données seront collectées. L'introduction comprend une première tâche de familiarisation comprenant 4 questions, sans

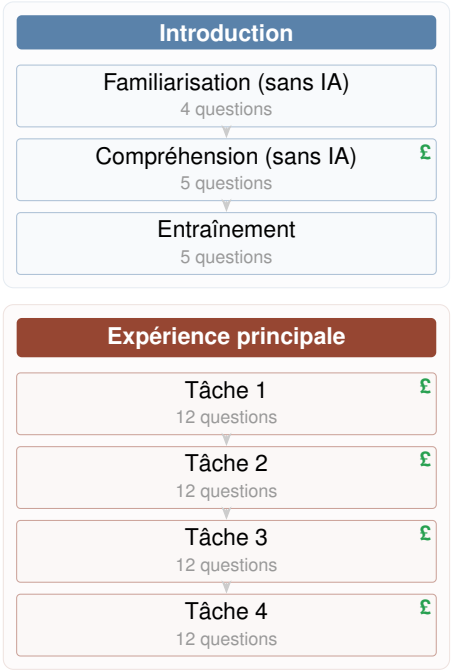


FIGURE 2 – Déroulement du protocole. Le symbole £ précise les tâches considérées pour la rémunération bonus.

limite de temps de réponse et sans assistance de l'IA, quel que soit le groupe. Elle comprend ensuite une tâche de compréhension (5 questions, sans limite de temps et sans assistance de l'IA), qui sera utilisée pour exclure de l'analyse statistique les participants n'ayant pas démontré une compréhension suffisante de la tâche. La tâche de compréhension est la seule tâche de l'introduction qui est prise en compte pour la rémunération bonus des participants, afin de les inciter à répondre du mieux possible. L'introduction se clôt avec une tâche d'entraînement comprenant 5 questions et dans laquelle l'assistance IA et ses éventuelles explications sont introduites pour les groupes concernés. Une limite de temps de réponse de 45 secondes est également introduite pour cette tâche. Les grilles de l'introduction sont de taille 12×12 (difficulté intermédiaire entre Facile et Difficile), sauf pour le test de compréhension pour lequel les grilles sont de taille 5×5. À l'issue de cette phase débute la phase expérimentale proprement dite. Elle se compose de quatre tâches de 12 questions chacune. Chaque tâche correspond à une combinaison unique des facteurs pression temporelle et difficulté de tâche, avec Pression $\in \{\text{Faible}, \text{Élevée}\}$ et Difficulté $\in \{\text{Facile}, \text{Difficile}\}$. L'ensemble des conditions expérimentales est ainsi $\{\text{Faible}, \text{Élevée}\} \times \{\text{Facile}, \text{Difficile}\}$, et l'ordre de présentation de ces conditions est tiré aléatoirement pour chaque participant. Chaque condition est associée à un motif spécifique et à un ensemble prédéterminé de 12 grilles. L'ordre de présentation des grilles pour chaque participant est déterminé aléatoirement. Pour chaque ensemble de 12 grilles, le modèle prédictif commet exactement deux erreurs (un faux positif et un faux négatif), identiques pour l'ensemble des participants. Cela correspond donc à un système d'assistance présentant une *accuracy* très proche de 85%. À l'issue de chaque tâche de l'expérience principale, les participants répondent à des questionnaires évaluant leur confiance et leur charge de travail. Leur score pour la tâche qu'ils viennent d'effectuer leur est ensuite communiqué. L'expérience principale inclut également deux tests d'attention[3], placés après la première et après la quatrième tâche. L'expérience a été conçue pour être réalisée de manière autonome, en ligne via le Web, pour une durée totale d'environ 20 minutes.

3.1.4 Participants

L'obtention des résultats a mobilisé 201 participants recrutés via Prolific[4]. Après exclusions pour problèmes techniques (5), échecs à un test d'attention (16) ou au test de compréhension (8), la cohorte retenue comprend 172 participants : 85 dans le groupe H et 87 dans le groupe H+IA. Tous sont adultes résidant au Royaume-Uni (55,8% de femmes, âge moyen de 44,0 ans). Les personnes présentant des troubles de vision des couleurs étaient exclues dès le recrutement, la tâche reposant sur la discrimination de symboles colorés. Les participants ont été rémunérés à un taux fixe de 6£/h, complété par un bonus de 0,04£ par bonne réponse aux 53 questions évaluées (soit un taux horaire effectif entre 6£ et 12,12£). L'assignation aux groupes H ou H+IA est aléatoire.

3.2 Modèles d'IA

La tâche de prédiction est une classification binaire (présence ou absence du motif) à partir d'une image RGB ($256 \times 256 \times 3$). Un modèle ResNet-18 [15] est entraîné de manière supervisée pour chaque couple (dimension de la grille, motif), à l'aide de jeux de données synthétiques (49k images d'entraînement, 1k de test). L'architecture atteignant aisément 100% de précision sur cette tâche, l'entraînement est volontairement interrompu dès que le seuil de 85% est atteint sur le jeu de test, afin de simuler un système d'assistance imparfait. Toutes les données incluses dans le protocole sont issues du jeu de test.

3.3 Implémentation et hébergement

Le protocole a été déployé sur une instance de la plateforme web WebXAII [19], hébergée sur des serveurs d'IMT Mines Alès. L'ensemble des pages nécessaires à l'expérience (informations, instructions, tâches de décision et questionnaires) y ont été implémentées.

3.4 Analyses statistiques

Deux analyses complémentaires sont menées pour étudier l'influence de l'ordre des tâches sur les variables dépendantes. Les hypothèses de normalité et d'homoscédasticité n'étant pas garanties avec des sous-groupes de petite taille, des tests non paramétriques U de Mann-Whitney bilatéraux

3. Chaque test comporte deux questions : l'une demande de sélectionner une réponse précise sur une échelle de Likert (ex. « Somewhat Disagree ») ; l'autre propose un choix parmi trois options, leur demandant la couleur du ciel ou le principe des tâches qu'ils réalisent.

4. https://www.prolific.com/

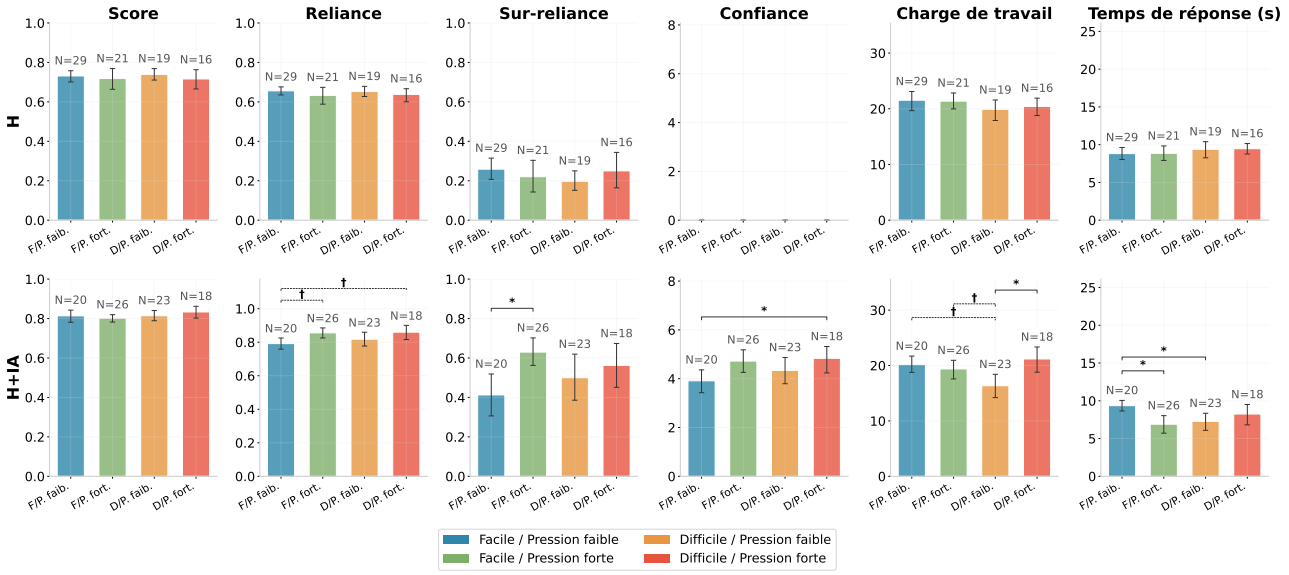


FIGURE 3 – Valeurs des variables dépendantes agrégées sur l'ensemble des tâches pour les conditions H et H+IA, avec intervalles de confiance à 95% obtenus par bootstrap, en fonction de la tâche initiale (combinaison difficulté × pression temporelle). Les différences significatives sont indiquées par des astérisques : * pour $p_{\text{BH}} < .05$, ** pour $p_{\text{BH}} < .01$ et *** pour $p_{\text{BH}} < .001$. Les différences marginales ($p_{\text{BH}} < .10$) sont signalées par le symbole †.

sont utilisés pour l'ensemble des comparaisons. La taille d'effet est estimée par le d de Cohen et les intervalles de confiance à 95% sont obtenus par bootstrap (10 000 rééchantillonnages). Les p-values sont corrigées par la procédure de Benjamini-Hochberg (BH-FDR), qui contrôle le taux de fausses découvertes tout en offrant une puissance supérieure à des corrections plus conservatrices telles que Bonferroni.

La première analyse porte sur l'effet de la tâche initiale. Les participants sont répartis en quatre groupes selon la première tâche effectuée, et les variables dépendantes sont agrégées sur l'ensemble des quatre tâches. Des tests de Mann-Whitney sont réalisés pour chacune des six paires de groupes, séparément par condition expérimentale (H, H+IA) et par variable dépendante, cette combinaison définissant la famille de correction BH-FDR.

La seconde analyse porte sur l'effet de l'exposition antérieure à un type de tâche donné, sur la condition H+IA uniquement. Pour chaque dimension (pression temporelle, difficulté) et chaque direction d'exposition, les observations sont réparties en trois groupes selon l'historique des tâches précédentes : aucune exposition au type étudié, exposition aux deux types, ou exposition uniquement au type étudié. Par exemple, pour l'analyse de l'effet d'un antécédent à pression forte, les trois groupes correspondent aux observations sans pression forte antérieure, avec pression forte et faible antérieures, et avec uniquement pression forte antérieure. Quatre séries de comparaisons sont ainsi effectuées, et les trois tests par paires au sein de chaque combinaison série × variable dépendante constituent la famille de correction. Cette analyse vise à tester si l'exposition antérieure à une condition exigeante a un impact persistant sur le comportement lors des tâches suivantes, en lien avec l'observation par Swaroop et al. selon laquelle la première exposition à une forte pression temporelle déclencherait un comportement de sur-reliance durable [27]. Nous cherchons à vérifier cet effet sur nos données et à déterminer si un effet similaire apparaît avec la première exposition à une tâche difficile.

Nous effectuons finalement une analyse des effets directs du genre et de l'âge sur les variables dépendantes, dans les deux groupes H et H+IA, respectivement par un test de Mann-Whitney bilatéral et une corrélation de Spearman, avec correction BH-FDR appliquée séparément pour chaque famille de tests (six tests par famille).

4 Résultats et discussions

Dans cette section, nous présentons les principaux résultats issus de l'expérimentation. Nous analysons dans un premier temps l'influence de la première tâche expérimentale sur le comportement global des participants, puis nous examinons plus finement les effets de la première exposition à une condition exigeante (forte pression temporelle ou difficulté élevée) ou peu exigeante (faible pression temporelle ou difficulté faible) sur le comportement lors des tâches ultérieures.

4.1 Effets de la tâche initiale sur le comportement global

La Figure 3 présente les variables dépendantes agrégées sur l'ensemble des quatre tâches, en fonction de la première condition expérimentale rencontrée par chaque participant, pour les groupes H (sans assistance) et H+IA. La Table 1 détaille les comparaisons par paires dont la p-value est inférieure à 0,1 – ce qui inclut l'ensemble des comparaisons

TABLE 1 – Comparaisons par paires selon la première condition ($p_{\text{BH}} < .10$).

Cond.	Variable	Comparaison	M_1	M_2	d	p_{MW}	p_{BH}
H+IA	Charge de travail	Diff./Press. faible vs Diff./Press. forte	16.348	21.181	-0.93	0.0058	**0.0346**
H+IA	Charge de travail	Fac./Press. faible vs Diff./Press. faible	20.200	16.348	+0.86	0.0213	0.0639
H+IA	Charge de travail	Fac./Press. forte vs Diff./Press. faible	19.385	16.348	+0.62	0.0495	0.0990
H+IA	Confiance	Fac./Press. faible vs Diff./Press. forte	3.913	4.833	-0.78	0.0029	**0.0173**
H+IA	Reliance	Fac./Press. faible vs Fac./Press. forte	0.792	0.856	-0.81	0.0100	0.0546
H+IA	Reliance	Fac./Press. faible vs Diff./Press. forte	0.792	0.860	-0.80	0.0182	0.0546
H+IA	Sur-reliance	Fac./Press. faible vs Fac./Press. forte	0.412	0.630	-1.01	0.0057	**0.0341**
H+IA	Temps de réponse	Fac./Press. faible vs Fac./Press. forte	9.353	6.894	+0.97	0.0071	**0.0247**
H+IA	Temps de réponse	Fac./Press. faible vs Diff./Press. faible	9.353	7.278	+0.89	0.0082	**0.0247**

significatives au seuil de 0,05.

Un premier résultat notable concerne le groupe H : aucune différence significative n'est observée entre les quatre groupes définis par la tâche initiale, que ce soit en termes de score, de charge de travail ou de temps de réponse. L'absence d'effet d'ordre dans la condition sans assistance constitue un résultat de contrôle important : il indique que les effets observés dans la condition H+IA ne sont pas imputables à des différences intrinsèques entre les groupes de participants ou à un simple effet d'apprentissage de la tâche, mais relèvent bien d'une interaction avec le système prédictif. Pour le groupe H+IA, les résultats révèlent des différences significatives en fonction de la tâche initiale. Les participants ayant débuté par la condition Facile/Pression faible se distinguent des autres sur plusieurs variables. Leur sur-reliance est significativement plus faible que celle des participants ayant débuté par Facile/Pression forte ($d = -1{,}01$, $p_{\text{BH}} = {,}034$). Leur confiance dans le modèle est significativement inférieure à celle des participants ayant débuté par Difficile/Pression forte ($d = -0{,}78$, $p_{\text{BH}} = {,}017$). Enfin, leurs temps de réponse sont significativement plus longs que ceux des participants ayant débuté par Facile/Pression forte ($d = +0{,}97$, $p_{\text{BH}} = {,}025$) ou par Difficile/Pression faible ($d = +0{,}89$, $p_{\text{BH}} = {,}025$). On observe également une tendance marginale pour la reliance, qui est plus faible chez les participants ayant débuté par Facile/Pression faible par rapport à ceux ayant débuté par Facile/Pression forte ou Difficile/Pression forte. La charge de travail présente un profil différent : les participants ayant débuté par Difficile/Pression faible rapportent une charge significativement inférieure à celle des participants ayant débuté par Difficile/Pression forte ($d = -0{,}93$, $p_{\text{BH}} = {,}035$), avec des tendances marginales par rapport aux deux groupes ayant débuté par une tâche facile. Ces résultats suggèrent que débuter par une tâche facile sous faible pression temporelle offre aux participants l'opportunité de développer une stratégie de résolution autonome. Il est probablement plus facile pour les participants de constater que le modèle prédictif commet des erreurs dans cette condition. Cette première expérience d'un travail effectif de résolution semble ancrer un comportement d'évaluation critique des prédictions qui perdure sur l'ensemble de l'expérience, se traduisant par une confiance et une sur-reliance plus faibles ainsi que des temps de réponse plus longs. À l'inverse, les participants confrontés d'emblée à une tâche exigeante semblent contraints de recourir rapidement à une stratégie de délégation au modèle, faute de ressources cognitives ou temporelles suffisantes pour résoudre la tâche de manière autonome. Toutefois, ces conclusions doivent être nuancées : les effets ne sont pas systématiques sur l'ensemble des combinaisons de facteurs. Par exemple, la comparaison entre les participants ayant débuté par Facile/Pression faible et ceux ayant débuté par Difficile/Pression forte n'atteint pas le seuil de significativité pour les temps de réponse, alors que les comparaisons impliquant un seul facteur sont significatives. Cela suggère que les mécanismes par lesquels la pression temporelle et la difficulté influencent le comportement ne sont pas strictement additifs.

4.2 Effets de l'exposition antérieure à une condition exigeante

La Figure 4 et la Table 2 détaillent les effets de l'exposition antérieure aux différentes conditions, en distinguant quatre analyses selon la dimension étudiée (pression temporelle ou difficulté) et la direction de l'effet. Les observations sont réparties en trois groupes selon l'historique des tâches précédentes : aucune exposition au type étudié, exposition aux deux types, ou exposition uniquement au type étudié.

4.2.1 Effets de l'exposition antérieure à une pression forte

Pour les observations réalisées sur des tâches à pression faible, les participants ayant déjà rencontré au moins une tâche à pression forte présentent une confiance significativement plus élevée que ceux n'ayant jamais été exposés à la pression forte. Cet effet est particulièrement marqué pour le groupe n'ayant connu que de la pression forte auparavant ($d = -0{,}80$, $p_{\text{BH}} = {,}001$) et reste significatif lorsque les deux types de pression ont été rencontrés ($d = -0{,}56$, $p_{\text{BH}} = {,}003$). Une tendance marginale dans la même direction est observée pour la sur-reliance. Les temps de réponse sont significativement plus courts pour les participants ayant été exposés aux deux types de pression ($d = +0{,}57$, $p_{\text{BH}} = {,}007$), avec une tendance marginale pour ceux n'ayant connu que la pression forte. Ces résultats indiquent que l'exposition préalable à une pression temporelle forte déclenche un comportement de délégation accrue

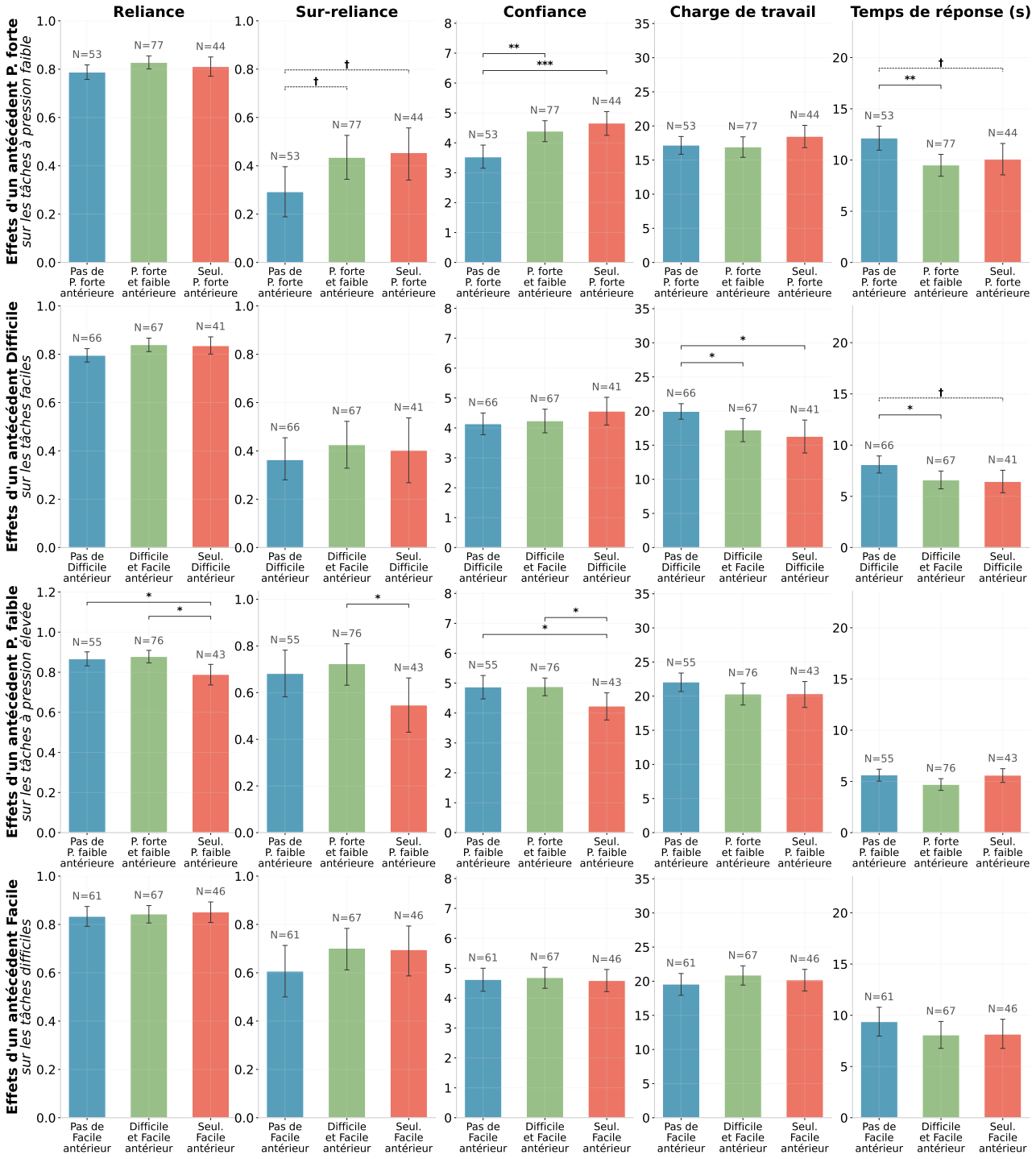


FIGURE 4 – Analyse des variables selon l'exposition antérieure aux conditions opposées, avec intervalles de confiance à 95% obtenus par bootstrap. Pour chaque type de tâche, les observations sont réparties en trois groupes selon l'historique des tâches précédentes : aucune exposition au type opposé, exposition aux deux types, ou exposition uniquement au type opposé. Les comparaisons par paires sont corrigées par famille par la procédure de Benjamini-Hochberg (BH-FDR). Les différences significatives sont indiquées par des astérisques : * pour $p_{\text{BH}} < .05$, ** pour $p_{\text{BH}} < .01$ et *** pour $p_{\text{BH}} < .001$. Les différences marginales ($p_{\text{BH}} < .10$) sont signalées par le symbole †.

TABLE 2 – Effet de l'exposition antérieure — Mann-Whitney U, correction BH-FDR, $p_{\text{BH}} < .10$ — H+IA

Analyse	VD	Comparaison	M_i	M_j	d	p_{BH}
Tâches à pression faible	Sur-reliance	Pas de P. forte ant. vs P. forte et faible ant.	0.292	0.435	−0.36	0.0640
		Pas de P. forte ant. vs Seul. P. forte ant.	0.292	0.455	−0.43	0.0640
	Confiance	Pas de P. forte ant. vs Seul. P. forte ant.	3.528	4.659	−0.80	**0.0005**
		Pas de P. forte ant. vs P. forte et faible ant.	3.528	4.390	−0.56	**0.0029**
	Tps réponse (s)	Pas de P. forte ant. vs P. forte et faible ant.	12.149	9.512	+0.57	**0.0072**
		Pas de P. forte ant. vs Seul. P. forte ant.	12.149	10.093	+0.43	0.0589
Tâches faciles	Charge de travail	Pas de Diff. ant. vs Diff. et Facile ant.	19.939	17.224	+0.45	**0.0469**
		Pas de Diff. ant. vs Seul. Diff. ant.	19.939	16.293	+0.58	**0.0469**
	Tps réponse (s)	Pas de Diff. ant. vs Diff. et Facile ant.	8.094	6.592	+0.42	**0.0208**
		Pas de Diff. ant. vs Seul. Diff. ant.	8.094	6.443	+0.47	0.0705
Tâches à pression forte	Reliance	P. forte et faible ant. vs Seul. P. faible ant.	0.878	0.789	+0.59	**0.0114**
		Pas de P. faible ant. vs Seul. P. faible ant.	0.867	0.789	+0.51	**0.0391**
	Sur-reliance	P. forte et faible ant. vs Seul. P. faible ant.	0.724	0.547	+0.44	**0.0418**
	Confiance	Pas de P. faible ant. vs Seul. P. faible ant.	4.873	4.233	+0.42	**0.0376**
		P. forte et faible ant. vs Seul. P. faible ant.	4.882	4.233	+0.46	**0.0376**

au modèle prédictif — se traduisant par une confiance plus élevée et des temps de réponse plus courts — qui persiste même lorsque la pression temporelle diminue. Ce résultat est cohérent avec les observations de Swaroop et al. [27], ce qui tend à confirmer l'existence d'un effet d'ancrage lié à la pression temporelle.

4.2.2 Effets de l'exposition antérieure à une tâche difficile

Pour les observations réalisées sur des tâches faciles, les participants ayant déjà rencontré au moins une tâche difficile rapportent une charge de travail significativement plus faible, aussi bien pour le groupe ayant connu les deux types de difficulté ($d = +0{,}45$, $p_{\text{BH}} = {,}047$) que pour celui n'ayant connu que des tâches difficiles ($d = +0{,}58$, $p_{\text{BH}} = {,}047$). Les temps de réponse sont également significativement plus courts pour le groupe ayant été exposé aux deux types ($d = +0{,}42$, $p_{\text{BH}} = {,}021$), avec une tendance marginale pour celui n'ayant connu que des tâches difficiles. L'exposition antérieure à la difficulté ne semble pas affecter directement la confiance ou la sur-reliance, contrairement à l'exposition à la pression temporelle. Son effet se manifeste plutôt par une réduction de la charge de travail perçue et des temps de décision. Ce résultat suggère que les participants ayant d'abord effectué des tâches difficiles tendent à évaluer les tâches faciles suivantes comme nécessitant moins d'effort, ce qui se traduit par des décisions plus rapides. Il pourrait s'agir d'un effet de contraste plutôt que d'un ancrage sur la reliance, les participants ajustant leur investissement cognitif à la baisse après avoir expérimenté un niveau de difficulté supérieur.

4.2.3 Effets de l'exposition antérieure à une pression faible et à une tâche facile

L'examen de l'impact d'une exposition antérieure à des conditions peu exigeantes sur le comportement lors de tâches exigeantes complète les analyses précédentes. Pour les tâches à pression forte, les différences significatives apparaissent systématiquement en comparaison au groupe n'ayant connu que de la pression faible auparavant : ce groupe présente une reliance, une sur-reliance et une confiance significativement plus faibles que les participants ayant déjà rencontré de la pression forte (que ce soit exclusivement ou en combinaison avec de la pression faible). Ces résultats confirment que c'est bien la rencontre préalable avec une tâche à pression forte qui déclenche le changement de comportement : les participants exclusivement exposés à la pression faible maintiennent un usage plus modéré du système d'assistance, tandis que dès lors qu'une tâche à pression forte a été rencontrée, le comportement de délégation s'installe indépendamment de l'historique ultérieur. En revanche, pour les tâches difficiles précédées de tâches faciles, aucune différence significative n'est observée sur l'ensemble des variables étudiées. L'exposition antérieure à des tâches faciles ne semble donc pas modifier le comportement des participants lors de tâches difficiles.

4.2.4 Asymétrie des effets d'ancrage

L'ensemble de ces résultats met en évidence une asymétrie dans les effets d'ancrage liés aux conditions expérimentales. Les conditions exigeantes exercent un effet d'ancrage sur le comportement lors des tâches ultérieures moins exigeantes : la pression temporelle forte ancre une confiance et une reliance élevées qui persistent, tandis que la difficulté élevée ancre une réduction de la charge de travail perçue et des temps de réponse. L'analyse des effets en fonction des antécédents à pression ou à difficulté faible confirme cette asymétrie : pour la pression temporelle, c'est la rencontre avec une tâche à pression forte qui constitue le facteur déterminant du changement de comportement, puisque les différences significatives impliquent systématiquement le

groupe n'ayant connu que de la pression faible. Pour la difficulté, l'exposition antérieure à des tâches faciles n'exerce aucun effet détectable sur le comportement lors de tâches difficiles. Cette asymétrie suggère que les conditions exigeantes constituent des ancres cognitives plus puissantes que les conditions peu exigeantes. D'un point de vue pratique, ce résultat implique que l'ordre dans lequel les opérateurs sont exposés aux différentes conditions de travail avec un système d'assistance par IA n'est pas anodin : une exposition progressive, débutant par des conditions peu contraignantes, pourrait favoriser une calibration plus appropriée de la confiance et de la reliance.

4.3 Variabilité interindividuelle

Dans le groupe H+IA, le genre et l'âge ne sont associés à aucune variable comportementale de manière significative (tous les $p_{\text{BH}} > ,23$), à l'exception de la charge de travail reportée, qui est significativement plus élevée chez les femmes que chez les hommes ($M_{\text{femmes}} = 21{,}2$ vs $M_{\text{hommes}} = 17{,}2$, $d = +0{,}89$, $p_{\text{BH}} < ,001$). Il s'agit d'un effet qui a déjà été observé dans la littérature [9]. Dans le groupe H, le genre n'est associé à aucune variable dépendante (tous les $p_{\text{BH}} > ,44$). En revanche, l'âge prédit négativement le score ($\rho = -0{,}28$, $p_{\text{BH}} = ,024$), et positivement la charge de travail ($\rho = +0{,}30$, $p_{\text{BH}} = ,024$). Ces effets de l'âge, absents dans le groupe H+IA, suggèrent que l'assistance par IA pourrait atténuer l'influence de l'âge sur la performance et la charge de travail.

5 Limites

Notre étude présente plusieurs limites qu'il convient de mentionner. Premièrement, les résultats sont issus d'une analyse post-hoc de données obtenues pour un protocole ayant pour but initial d'examiner l'impact de techniques d'explicabilité de l'IA. L'assignation des participants aux conditions de tâche initiale résulte d'une randomisation de l'ordre des tâches, ce qui engendre des groupes de tailles inégales. Un protocole dédié, avec un contrôle explicite de la tâche initiale et des groupes plus équilibrés, serait nécessaire pour confirmer les effets observés avec une puissance statistique accrue.

Deuxièmement, la tâche proposée constitue une abstraction qui ne reflète pas directement les conditions réelles d'utilisation des systèmes d'aide à la décision. Si cette simplification permet un contrôle rigoureux de la difficulté et de la pression temporelle, elle limite la généralisation des résultats à des contextes professionnels spécifiques. Des travaux futurs pourraient reproduire ces analyses avec des tâches plus proches de situations réelles, dans des domaines où les systèmes d'aide à la décision par IA sont déjà déployés, tels que la radiologie ou le domaine militaire. L'expertise métier des opérateurs spécialisés de ces domaines pourrait moduler la résistance aux effets d'ancrage.

Troisièmement, le système d'assistance présente un taux d'erreur fixe d'environ 15%, identique pour toutes les conditions. Dans un contexte réel, les performances d'un modèle prédictif pourraient varier selon la difficulté des instances. Il serait intéressant, dans des travaux futurs, de faire varier indépendamment la difficulté des instances et le taux d'erreur du modèle, afin de mesurer la contribution propre de chacun de ces facteurs sur la reliance des opérateurs humains.

6 Conclusion

Cette étude contribue à montrer que la construction de la relation de confiance entre un opérateur humain et un système prédictif d'IA peut dépendre des premières interactions avec ce système. En analysant le comportement de 172 participants dans une tâche de détection de motifs assistée par IA, nous mettons en évidence des effets d'ancrage liés aux conditions de pression temporelle et de difficulté rencontrées lors des premières tâches expérimentales. Nos résultats montrent que les participants confrontés d'emblée à des conditions exigeantes adoptent rapidement une stratégie de délégation au modèle prédictif, qui persiste même lorsque les conditions deviennent plus favorables. À l'inverse, les participants débutant par une tâche facile sous faible pression tendent à avoir une reliance et sur-reliance plus faible sur l'ensemble de l'expérience. L'analyse des effets d'exposition antérieure révèle une asymétrie : les conditions exigeantes constituent des ancres cognitives plus puissantes, la pression temporelle agissant sur la confiance et la sur-reliance, la difficulté sur la charge de travail et les temps de décision. Ces résultats impliquent que l'ordre d'exposition aux conditions de travail avec un système d'assistance par IA n'est pas anodin. Une introduction progressive, débutant par des conditions peu contraignantes, pourrait permettre aux opérateurs de calibrer leur confiance à sa juste valeur avant d'être confrontés à des situations exigeantes. Cette observation pourrait être pertinente tant dans un cadre éducatif que pour le déploiement opérationnel de systèmes d'assistance par IA. Par ailleurs, une analyse de la variabilité interindividuelle montre que dans le groupe n'ayant pas accès à l'assistance de l'IA, l'âge est corrélé négativement à la performance et positivement à la charge de travail. Dans le groupe avec l'assistance (H+IA), ces effets ne sont pas observés, ce qui suggère que l'IA permet de gommer une baisse de performance liée à l'âge. En revanche, nous observons dans le groupe ayant accès à l'assistance une charge de travail reportée plus élevée chez les femmes. Des travaux futurs devront confirmer ces résultats à l'aide de protocoles dédiés et de tâches plus représentatives de contextes métiers réels.

Remerciements

Ce travail est soutenu par le programme de recherche et d'innovation HORIZON de l'Union européenne, convention de subvention n° 101120657, projet ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI). Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources AD011011309R5 attribuée par GENCI.

Références

[1] Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. How AI Ideas Affect the Creativity, Diversity, and Evolution of Human Ideas : Evidence From a Large, Dynamic Experiment. In *Proceedings of the ACM Collective Intelligence Conference*, CI '25, pages 198–213, New York, NY, USA, August 2025. Association for Computing Machinery.

[2] Shreyan Biswas, Alexander Erlei, and Ujwal Gadiraju. Belief Updating and Delegation in Multi-Task Human-AI Interaction : Evidence from Controlled Simulations, February 2026. arXiv :2602.01986 [cs].

[3] R. J. M. Bruls and R. M. Kwee. Workload for radiologists during on-call hours : dramatic increase in the past 15 years. *Insights into Imaging*, 11(1) :121, November 2020.

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To Trust or to Think : Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1) :1–21, April 2021. arXiv :2102.09692 [cs].

[5] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2) :277 :1–277 :26, October 2023.

[6] Lemuria Carter and Dapeng Liu. How was my performance ? Exploring the role of anchoring bias in AI-assisted decision making. *International Journal of Information Management*, 82 :102875, June 2025.

[7] Tirtha Chanda, Katja Hauser, Sarah Hobelsberger, Tabea-Clara Bucher, Carina Nogueira Garcia, Christoph Wies, Harald Kittler, Philipp Tschandl, Cristian Navarrete-Dechent, Sebastian Podlipnik, Emmanouil Chousakos, Iva Crnaric, Jovana Majstorovic, Linda Alhajwan, Tanya Foreman, Sandra Peternel, Sergei Sarap, İrem Özdemir, Raymond L. Barnhill, Mar Llamas-Velasco, Gabriela Poch, Sören Korsing, Wiebke Sondermann, Frank Friedrich Gellrich, Markus V. Heppt, Michael Erdmann, Sebastian Haferkamp, Konstantin Drexler, Matthias Goebeler, Bastian Schilling, Jochen S. Utikal, Kamran Ghoreschi, Stefan Fröhling, Eva Krieghoff-Henning, and Titus J. Brinker. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nature Communications*, 15(1) :524, January 2024.

[8] M L Cummings. Automation Bias in Intelligent Time Critical Decision Support Systems. *Decision making in aviation*, 2004.

[9] Mary Lynne Dittmar, Joel S. Warm, William N. Dember, and David F. Ricks. Sex Differences in Vigilance Performance and Perceived Workload. *The Journal of General Psychology*, 120(3) :309–322, July 1993. _eprint : https ://doi.org/10.1080/00221309.1993.9711150.

[10] Nuša Farič, Sue Hinder, Robin Williams, Rishi Ramaesh, Miguel O Bernabeu, Edwin Van Beek, and Kathrin Cresswell. Early experiences of integrating an artificial intelligence-based diagnostic decision support system into radiology settings : a qualitative study. *Journal of the American Medical Informatics Association*, 31(1) :24–34, December 2023.

[11] Alicia Freel, Sabid Bin Habib Pias, Selma Šabanović, and Apu Kapadia. How Misclassification Severity and Timing Influence User Trust in AI Image Classification : User Perceptions of High- and Low-Stakes Contexts. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 2906–2923, New York, NY, USA, June 2025. Association for Computing Machinery.

[12] Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1) :35–42, February 2011.

[13] Nikita Haduong and Noah A. Smith. How Performance Pressure Influences AI-Assisted Decision Making, February 2025. arXiv :2410.16560 [cs].

[14] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index) : Results of Empirical and Theoretical Research. In Peter A. Hancock and Najmedin Meshkati, editors, *Advances in Psychology*, volume 52 of *Human Mental Workload*, pages 139–183. North-Holland, January 1988.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[16] Lukas Hermanns and Timm Teubner. Under pressure : how time constraints, task complexity, and AI reliability shape human-AI interaction. *Behaviour & Information Technology*, 0(0) :1–25, December 2025. _eprint : https ://doi.org/10.1080/0144929X.2025.2587732.

[17] Lujain Ibrahim, Katherine M. Collins, Sunnie S. Y. Kim, Anka Reuel, Max Lamparth, Kevin Feng, Lama Ahmad, Prajna Soni, Alia El Kattan, Merlin Stein, Siddharth Swaroop, Ilia Sucholutsky, Andrew Strait, Q. Vera Liao, and Umang Bhatt. Measuring and mitigating overreliance is necessary for building human-compatible AI, September 2025. arXiv :2509.08010 [cs].

[18] John D. Lee and Katrina A. See. Trust in Automation : Designing for Appropriate Reliance. *Human Factors*, 46(1) :50–80, March 2004.

[19] Jules Leguy, Pierre-Antoine Jean, Felipe Torres Figueroa, and Sébastien Harispe. WebXAII : an open-source web framework to study human-XAI interaction, May 2025.

[20] Michael Mayer. Trusting machine intelligence : artificial intelligence and human-autonomy teaming in military operations. *Defense & Security Analysis*, 39(4) :521–538, October 2023. _eprint : https ://doi.org/10.1080/14751798.2023.2264070.

[21] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, pages 340–350, New York, NY, USA, April 2021. Association for Computing Machinery.

[22] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding Fast and Slow : The Role of Cognitive Biases in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1) :83 :1–83 :22, April 2022.

[23] Stephen Rice, David Keller, Gayle Hunt, and David Trafimow. Automation Dependency Under Time Pressure. *2009 International Symposium on Aviation Psychology*, pages 611–616, January 2009.

[24] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your AI : expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 240–251, Marina del Ray California, March 2019. ACM.

[25] Nicolas Scharowski, Sebastian A. C. Perrig, Nick von Felten, and Florian Brühlmann. Trust and Reliance in XAI – Distinguishing Between Attitudinal and Behavioral Measures, March 2022. arXiv :2203.12318 [cs].

[26] Sonia Jawaid Shaikh and Ignacio F. Cruz. AI in human teams : effects on technology use, members' interactions, and creative performance under time scarcity. *AI & SOCIETY*, 38(4) :1587–1600, August 2023.

[27] Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, pages 138–154, New York, NY, USA, April 2024. Association for Computing Machinery.

[28] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. The effects of explanations on automation bias. *Artificial Intelligence*, 322 :103952, September 2023.

[29] Ruoxin Yang, Sisheng Li, Yawei Qi, Jiali Liu, Qinghua He, and Haichao Zhao. Unveiling users' algorithm trust : The role of task objectivity, time pressure, and cognitive load. *Computers in Human Behavior Reports*, 18 :100667, May 2025.

[30] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 295–305, New York, NY, USA, January 2020. Association for Computing Machinery.

[31] Zelun Tony Zhang, Felicitas Buchner, Yuanting Liu, and Andreas Butz. You Can Only Verify When You Know the Answer : Feature-Based Explanations Reduce Overreliance on AI for Easy Decisions, but Not for Hard Ones. In *Proceedings of Mensch und Computer 2024*, pages 156–170, Karlsruhe Germany, September 2024. ACM.

Évaluer la sobriété d'un service numérique : proposition d'une grille d'évaluation multicritère adaptée au « *fuzzy front end* »

Robert Viseur[1]

1 UMONS, Faculté Warocqué d'Économie et de Gestion

robert.viseur@umons.ac.be

Résumé

La digitalisation est moins une dématérialisation qu'un transfert de support : terminaux, réseaux, datacenters, fabrication et fin de vie portent une part « invisible » des impacts. Nous soutenons que l'enjeu n'est pas seulement d'optimiser, mais de définir ce qui mérite d'être produit et consommé. Nous proposons une grille de sobriété numérique (consommation et production) et discutons le cas des services d'IA générative.

Mots-clés

Sobriété numérique, Green IT, Green IS, IA générative.

Abstract

Digitalisation is less dematerialisation than a transfer of support: devices, networks, data centres, manufacturing and end-of-life account for much of the 'invisible' footprint. We argue that the key challenge is not only optimisation, but deciding what deserves to be produced and consumed. We propose a digital sobriety scoring framework (use and supply) and discuss the specific pressures introduced by generative AI services.

Keywords

Digital sobriety, Green IT, Green IS, GenAI.

1 Introduction

En 2021, le nombre de terminaux connectés s'élevait à 27,1 milliards d'unités [13], à comparer aux 13,4 milliards relevés en 2016, soit une augmentation de 102,24 % sur cinq ans. Dans cet ensemble, les *smartphones* représentent près de 25 % du matériel tandis que les objets connectés comptent pour plus de la moitié [13]. Bordage voit dès lors la digitalisation moins comme une dématérialisation que comme une forme de « *transfert de support* » [8]. Berthoud et Parry évoquent pour leur part des « *transferts de pollution* » (p. 16) [6]. Cette matérialité des TIC passe par les machines détenues par les particuliers et les organisations mais aussi celles associées aux services en ligne sur lesquels elles se connectent. Les services numériques questionnent donc quant à leur impact sur les écosystèmes naturels. Cette recherche, dédiée à l'évaluation de la sobriété d'un service numérique, répond dès lors à l'appel d'universitaires pour positionner les sciences de gestion dans l'ère de l'anthropocène [1]. S'inscrivant dans un naturalisme aménagé [1], elle propose, premièrement, de documenter les impacts écologiques des services numériques, deuxièmement, de structurer une grille d'évaluation de la sobriété d'un service numérique, troisièmement, de discuter plus spécifiquement les impacts liés aux services d'intelligence artificielle (IA) générative.

2 Impact environnemental des équipements informatiques

Dans cette section, nous introduisons la problématique des impacts du numérique. Nous traitons ensuite de leur réduction et, enfin, de leur mesure.

2.1 Ordres de grandeur

Dans son rapport intitulé « *Déployer la sobriété numérique* », le *think tank* The Shift Project présente une ventilation de la consommation d'énergie finale du numérique [19]. La consommation d'énergie se répartit globalement entre l'utilisation (55 %) et la production (45 %) des équipements (cf. Tableau 1). Ces chiffres appellent quelques constatations. Premièrement, la consommation à l'utilisation se répartit de manière relativement équilibrée entre les terminaux, les réseaux et les centres de données. Près des deux tiers de la consommation énergétique, associée au traitement des données et à leur transfert, est donc peu ou prou visible pour l'utilisateur. Elle doit cependant être comptabilisée dans l'impact environnemental lié à l'utilisation d'un service numérique. Deuxièmement, près de la moitié de l'impact environnemental est déterminé par la production des équipements. Troisièmement, ces chiffres n'incluent pas la consommation associée au traitement en fin de vie.

Tableau 1. Consommation d'énergie par poste.

Utilisation	Terminaux	20 %	55 %	36,36 %
	Réseaux	16 %		29,09 %
	Datacentres	19 %		34,55 %
Production	Ordinateurs	17 %	45 %	37,78 %
	Smartphones	11 %		24,44 %
	Télévisions	11 %		24,44 %
	Autres	6 %		13,33 %

Ces chiffres se concentrent sur la consommation énergétique, qui ne reflète cependant qu'une partie de l'impact environnemental du numérique [17]. La production s'accompagne ainsi d'une consommation de

métaux, d'eau ainsi que de rejets de déchets et de gaz à effet de serre (GES). Quant au traitement en fin de vie, il se caractérise par un recyclage limité, entraînant la production de déchets, parfois exportés, ainsi que des rejets de polluants. Par ailleurs, ces équipements, et en particulier les *smartphones* [53], pèsent lourd dans la consommation de certains métaux. Or, pour la grande majorité des métaux, les réserves estimées, parfois géographiquement concentrées, se situent entre 30 et 60 ans [3]. De plus, le recyclage de ces métaux s'avère faible pour les biens numériques, car leur séparation est souvent difficile.

2.2 Réduction de l'impact environnemental

En pratique, la réduction de l'impact environnemental du numérique passe classiquement par deux mesures, d'une part, l'allongement de la durée de vie des équipements, d'autre part, l'optimisation de l'infrastructure technique [54]. La première concerne tant les consommateurs que les organisations, privées ou publiques, tandis que la seconde est essentiellement l'apanage des organisations. Ces dernières sont encouragées à réduire leur impact environnemental dès lors qu'elles supportent directement la consommation de ressources. Dans le cas où l'obésité logicielle est essentiellement supportée par l'utilisateur (cas des systèmes d'exploitation ou des suites bureautiques par exemple), l'incitation à optimiser pesant sur l'entreprise éditrice est cependant réduite, ce qui suppose une démarche volontaire de la part de celle-ci. Sur un plan financier, la gestion du Coût Total de Possession (TCO) des systèmes informatiques encourage l'achat de matériel fiable et réparable tandis que, du fait des cycles d'innovation (modèle A-U), les produits et services numériques matures tendent à privilégier l'innovation de processus, avec les économies de coûts de production associés. Dans ce contexte, réduire les coûts s'accompagne souvent d'une moindre consommation de ressources [47].

La lutte contre l'obsolescence des terminaux représente un enjeu important compte tenu de l'impact lié à la fabrication [8, 54]. Les types de machines doivent être distingués. Ainsi, sur le plan des émissions de CO_2, la fabrication d'un *smartphone*, d'un ordinateur portable et d'un serveur compte respectivement pour 72 %, 52 % et 20 % des émissions sur le cycle de vie de l'appareil [54]. Ces différences s'expliquent notamment par le taux d'utilisation des équipements, plus élevé pour les serveurs du fait notamment de l'usage de technologies comme la virtualisation. Les mesures favorables à l'allongement de la durée de vie du matériel sont cependant connues [38]. Elles incluent le réemploi (réparation et réutilisation de l'appareil ; p. ex. Back Market et ITJustGood anciennement Tradediscount), la réutilisation (réutilisation de sous-ensembles) et le recyclage (réutilisation des matières premières ; p. ex. Fairphone). La réparabilité, incluant une plus grande modularité des appareils [8], permet par ailleurs de faciliter le réemploi. Elle peut être favorisée par l'allongement des durées de garantie (mesure légale), mais aussi par une meilleure information au consommateur (voir, par exemple, les indices de réparabilité portés par iFixIt ou Spareka).

L'optimisation de l'infrastructure technique s'appuie sur quatre leviers [42]. Le premier est le logiciel, incluant en particulier le modèle sous-jacent dans le cas des systèmes d'intelligence artificielle. Le second est le matériel permettant son exécution. Le troisième est le centre de données hébergeant le matériel. Le quatrième concerne les sources d'énergie utilisées par le centre de données. Ainsi, les entreprises gérant les grands centres de données veillent à décarboner l'alimentation en électricité, soit par leurs propres capacités de production, soit par la souscription de contrats d'approvisionnement d'électricité verte [29]. Les centres de données peuvent être, d'une part, localisés dans des zones froides, d'autre part, conçus pour abaisser les coûts de refroidissement. Les équipements informatiques peuvent être adaptés aux usages (p. ex. élimination des composants inutiles), et des puces spécialisées utilisées sur certaines tâches [55]. Enfin, les logiciels exécutés dans ces centres peuvent être optimisés, par le choix d'algorithmes et de langages de programmation plus efficaces [55].

Deux exemples permettent d'illustrer ces stratégies d'optimisation. Le premier cas, celui de Meta (avec Facebook), se traduit par l'innovation collaborative sur la conception des centres de données au sein de la plateforme Open Compute Project. Créée en 2011, elle se définit comme « *une communauté collaborative axée sur la refonte de la technologie matérielle afin de répondre efficacement aux demandes croissantes en matière d'infrastructure informatique* ». Meta revendique une économie d'énergie de l'ordre de 38 % à la construction et une réduction du coût de fonctionnement de l'ordre de 24 %. De plus, après avoir utilisé PHP, un langage interprété, Meta a développé un traducteur de code source vers le C++, un langage compilé (projet open-source « HipHop for PHP »). Les limites de cette approche ayant été atteintes, malgré une réduction de 50 % de la charge processeur, un nouveau langage de programmation hybride baptisé Hack (projet open-source « HHVM ») a finalement été lancé. Le second cas, celui de Netflix, s'appuie plutôt sur une externalisation de son infrastructure auprès d'une entreprise d'hébergement dont c'est la spécialité (Amazon Web Services). Netflix concentre par contre ses ressources sur son cœur de métier (le *streaming* vidéo) et n'hésite pas à ouvrir sa R&D à d'autres partenaires. Publié en 2018, le codec AV1 revendique ainsi un gain de bande passante de l'ordre de 30 % [27]. En outre, Netflix collabore avec les FAI via son programme Open Connect. Cette politique s'inscrit par ailleurs dans le cadre d'un accord entre Netflix et le régulateur européen visant à modérer les volumes de données transférés [21].

2.3 Comprendre et mesurer l'impact environnemental des *datacenters*

Au-delà de l'empreinte écologique à la fabrication, les centres de données consomment généralement deux types de ressources : l'électricité et l'eau. Le refroidissement de machines peut recourir à différentes techniques [18]. Le refroidissement par détente directe, plus énergivore, recourant à la climatisation (groupe frigorifique), est généralement réservé aux petites unités modulaires. Le refroidissement gratuit (*free cooling*), très répandu, et moins énergivore, recoure, du plus économe au moins

économe en eau, soit à de l'air froid (rafraîchissement direct), soit à de l'air humidifié (rafraîchissement adiabatique), soit à de l'eau (rafraîchissement induit). Les développements récents se portent sur le refroidissement liquide direct où les processeurs sont par exemple immergés dans un fluide caloporteur. Dans le cas du rafraîchissement induit, l'impact environnemental dépend également de la provenance de l'eau. S'agit-il d'eau potable, d'eau usée ou d'eau de surface ? Dans le cas de Google, près de 80 % de l'eau prélevée est de l'eau de distribution (potable) mais la *bigtech* possède également des centres de données utilisant les eaux usées (Douglas, USA), l'eau de mer (Hamina, Finlande) ou l'eau de canal industriel (Saint-Ghislain, Belgique) (Google, 2025). Cette eau est ensuite rejetée (à une température supérieure) ou consommée (c'est-à-dire évaporée et réintroduite dans le grand cycle de l'eau [9]). La localisation du centre de données joue donc sur son impact. Premièrement, la présence d'un important réservoir (et d'un exutoire) d'eau (p. ex. mer) permet par exemple de le réduire. Globalement, la consommation des centres de données reste faible comparé à la consommation totale [40]. A contrario, la concentration des installations peut conduire à un impact local significatif, assorti de mesures visant à réduire la consommation. C'est par exemple le cas des centres situés le long de la Tamise à Londres [49]. Plus globalement, 14 % des prélèvements des centres de données de Google se déroulent dans des zones présentant un risque élevé de stress hydrique [22]. Deuxièmement, en fonction du mix énergétique, la disponibilité d'énergie décarbonée (énergie renouvelable, type solaire ou éolien, et nucléaire) permet d'en réduire sensiblement l'empreinte carbone. Là aussi, la concentration des grands centres de données peut venir perturber l'approvisionnement électrique. C'est notamment le cas en Irlande où des politiques d'atténuation des impacts ont été prises suite au succès des mesures incitatives à l'installation [49].

Dès lors, deux indicateurs principaux vont être suivis : le PUE et le WUE [18, 40]. Le PUE (*Power Usage Effectiveness*) est défini comme le rapport entre l'énergie totale annuelle consommée par le centre de données et l'énergie totale annuelle consommée par les seuls équipements IT. Le PUE moyen des centres de données industriels est de 1,58 contre 1,1 environ pour les fournisseurs d'infrastructures publiques de *cloud computing* [42]. La réduction du PUE passe notamment par le *free cooling*, soit que le centre de données soit refroidi naturellement par air froid, soit que l'eau soit utilisée comme fluide caloporteur. Le WUE (*Water Use Effectiveness*) est défini comme le ratio entre la quantité d'eau consommée par l'ensemble du centre de données et la consommation énergétique totale (L/kWh). Un indicateur supplémentaire peut être pris en compte pour évaluer la réutilisation de l'énergie. L'ERE (*Energy Reuse Effectiveness*) représente ainsi le rapport entre l'énergie totale consommée moins l'énergie réutilisée et l'énergie consommée par les équipements informatiques. Par exemple, Infomaniak a documenté son projet D4, un modèle ouvert de centre de données où la chaleur est réutilisée pour assurer du chauffage urbain. Son installation genevoise, qui vise une capacité de 10.000 serveurs, fournira un réseau de chaleur avec une capacité de 1,7MW, soit de quoi chauffer 6.000 logements (type Minergie-A) et éviter 3600 tCO_2eq chaque année. Au premier trimestre 2026, le PUE et le ERE de D4 tournaient respectivement autour de 1,1 et 0,25. Près de 90 % de l'énergie y était réutilisée tandis que le WUE se révélait très faible (pas d'utilisation d'eau additionnelle).

2.4 Mesurer les impacts environnementaux des services numériques

Plusieurs méthodes permettent de mesurer les impacts environnementaux d'un produit ou d'un service [6]. La première est le bilan carbone. Ce dernier ne se focalise par contre que sur un seul indicateur : les émissions de gaz à effet de serre. Le second est l'analyse économique intrants-sortants. Les flux monétaires sont alors analysés pour estimer les ressources et l'énergie utilisées par un produit ou un service, puis pour évaluer les émissions. Troisièmement, l'analyse de cycle de vie vise à évaluer tous les impacts environnementaux potentiels d'un produit, d'un service ou d'un procédé sur l'ensemble de son cycle de vie.

La comptabilisation au sein d'un bilan carbone distingue en pratique trois « scopes » (définis au sein de la norme ISO 14064). Le « scope 1 » concerne les émissions directes de gaz à effet de serre. Le « scope 2 » concerne les émissions indirectes liées à l'énergie tandis que le « scope 3 » concerne les émissions indirectes de l'entreprise liées à ses achats. Dans le cas des centres de données, on distinguera ainsi les émissions directes du centre de données (« scope 1 »), les émissions liées à la production électrique nécessaire pour le faire fonctionner (« scope 2 ») et les émissions indirectes (« scope 3 ») liées, par exemple, à la production et la fin de vie des équipements informatiques.

L'Analyse de Cycle de Vie (ACV) se déroule en quatre phases [6]. La première phase consiste à définir les objectifs et le champ de l'étude. En particulier, les frontières du système y sont déterminées. La seconde phase porte sur l'inventaire du cycle de vie, ce qui inclut le recensement et la quantification des flux de matière. La troisième phase concerne l'évaluation de l'impact du cycle de vie. Elle produit des indicateurs permettant de traduire les impacts du système considéré. La quatrième phase est dédiée à l'interprétation du cycle de vie. Elle veille à rendre les résultats compréhensibles et à documenter les limites de l'étude. L'évaluation nécessite la disponibilité de données statistiques, associées à « *leur lot d'incertitudes, d'approximations et d'hypothèses* » (p. 4) [6]. Rassemblées en bases de données (p. ex. EcoInvent), traitée par des logiciels spécialisés (p. ex. SimaPro), leur traitement conduit à des écarts parfois importants. La mise en œuvre de l'ACV nécessite donc du temps et des expertises pointues.

En informatique, l'optimisation renvoie à l'ensemble des techniques, méthodes et algorithmes conçus pour réduire l'utilisation des ressources tout en maintenant ou améliorant les performances du système. Elle consiste plus précisément à analyser puis ajuster le logiciel et l'environnement d'exécution afin d'éliminer le travail

inutile et d'améliorer les performances [25]. Les gains d'efficacité ne conduisent pas nécessairement à la réduction des usages dès lors qu'ils s'accompagnent d'un effet rebond [51]. La notion d'effet rebond trouve son origine dans le paradoxe de Jevons, selon lequel l'amélioration de l'efficacité d'une technologie conduit paradoxalement à une augmentation de la consommation de la ressource concernée plutôt qu'à sa réduction [23]. La littérature distingue l'effet rebond direct (la réduction de prix, provoquée par celle du coût, entraîne un accroissement de la demande pour le produit moins cher) et l'effet rebond indirect (l'économie faite sur un produit induit l'accroissement de la consommation d'un autre produit). Nous nous concentrerons dans cette recherche sur l'effet rebond direct. La frugalité numérique suppose l'atteinte des mêmes résultats avec moins d'énergie en étant plus économe en ressources informatiques [54]. L'on peut dresser un parallèle avec l'innovation frugale [30], qui cherche à atteindre les performances attendues par le marché, avec des produits moins complexes, centrés sur les fonctionnalités cœurs, à moindre coût, de manière à atteindre une clientèle précédemment non solvable. A contrario la sobriété numérique s'oppose à la croissance et à la (sur-)consommation. Elle vise ainsi à réduire l'usage des technologies numériques, d'une part, en se limitant à ce qui est réellement utile, d'autre part, en promouvant la longévité des équipements existants [8, 51]. Tant la frugalité que la sobriété numérique peuvent s'appuyer sur des *low-tech*. L'innovation *low-tech* vise à « *réduire la taille, l'intensité et la complexité du système technique (…) [de manière à entraîner] une diminution de l'empreinte environnementale d'un objet particulier* » (p. 179) [34]. Comparativement à la frugalité, l'atteinte de la sobriété suppose un effort [26] de manière à éviter la production de nouveaux produits et services. Elle nécessite donc aussi un changement de comportement de consommateur, en particulier pour éviter l'effet rebond [34].

3 Construire une grille d'évaluation de la sobriété numérique

Quelle pourrait être la « *sobriété par design* », au sens de Franck Aggeri [2], d'un service numérique ? Ce que nous proposons de construire n'est pas un substitut aux démarches d'ACV, mais une grille d'évaluation, basée sur une échelle de notation, destinée à outiller la discussion au sein des équipes produit dès les premières phases de l'innovation (c'est-à-dire le *fuzzy front end*). Cette proposition répond à un angle encore peu couvert par la littérature : disposer d'un instrument suffisamment simple pour interroger le sens de l'innovation [2], sans exiger d'emblée la complexité, les données et l'expertise associées à un chiffrage précis de type ACV [6]. De tels chiffrages s'exposent par ailleurs à l'indisponibilité des chiffres ou à l'incomplétude des méthodologies de mesure ([37] ; cf. page 15). Sur le plan des usages, les impacts se révèlent différenciés si l'on compare, par exemple, d'une part, le Bitcoin (compensation par conception des gains d'efficacité matérielle) et la publicité ciblée (inflation dans la collecte de données à caractère personnel, accélération des flux physiques), d'autre part, la vidéo en ligne et la visioconférence (réduction des déplacements ou des supports physiques) [55]. La littérature tend d'ailleurs à séparer les usages propices à la durabilité des optimisations des infrastructures numériques. Deux disciplines sont ainsi distinguées : le « Green IT » et le « Green IS ». Le terme « Green IT » renvoie à « *la réduction des effets environnementaux négatifs des technologies de l'information en utilisant et en éliminant les ressources informatiques de manière écoénergétique et rentable, en améliorant l'efficacité énergétique, en diminuant les émissions, ainsi qu'en réutilisant et en recyclant les matériaux* » tandis que le terme « Green IS » concerne « *la configuration et l'application des systèmes d'information afin d'atteindre des objectifs environnementaux, en réduisant l'empreinte écologique des entreprises et en soutenant la prise de décision organisationnelle vers la durabilité, tout en améliorant la performance économique* » (p. 938) [32]. Aligner durabilité (« *doing good* ») et profitabilité (« *doing well* ») pourrait ainsi relever du challenge pour les entreprises [58]. Nous distinguons dès lors deux types de critères, soit liés à la **consommation** (quel est l'objectif du service numérique ?), soit liés à la **production** (quelle est l'efficacité des infrastructures permettant au service de fonctionner ?).

3.1 Critères liés à la consommation

Premièrement, le numérique constitue un adjuvant pour la mise en œuvre d'une économie régénérative, non seulement capable de réduire les dommages sur les écosystèmes mais aussi de les restaurer et de les revitaliser [11]. Le numérique est ainsi vu comme un moyen de réduire l'empreinte environnementale des réseaux électriques, des transports, de l'agriculture ou de l'habitat [21]. Ces usages régénératifs coexistent cependant avec des usages extractivistes classiques tels que la publicité ciblée ou d'autres à visée récréative [55]. Un critère lié au **potentiel de régénération** peut dès lors être introduit. Dans quelle mesure le service rendu contribue-t-il aux durabilités environnementale, sociale et économique promues dans les Objectifs de Développement Durable ?

Deuxièmement, les services numériques présentent des conséquences en sens divers sur le plan des flux physiques induits par la technologie. Le numérique est souvent associé à une accélération des flux physiques. Peut ainsi être citée la publicité ciblée qui entraîne non seulement un traitement massif de données mais aussi une consommation accrue de biens physiques [55]. A contrario, certains usages relèvent d'une « *sobriété par le numérique* » [55], c'est-à-dire d'un changement de paradigme du physique vers le numérique [33]. Ainsi, et paradoxalement, la vidéo en ligne ou la visioconférence présentent, malgré la sollicitation des infrastructures numériques, une réduction substantielle des flux physiques. Idéalement, la tendance à la hausse ou la baisse doit cependant être estimée par ACV dès lors que toute forme de dématérialisation n'est pas nécessairement intéressante sur le plan des impacts environnementaux. Berthoud et Parry démontrent ainsi les limites de la dématérialisation dans les cas de l'affranchissement de colis (affranchissement traditionnel *vs* affranchissement

dématérialisé) et de la lecture de documents (feuilles papier *vs* clefs USB) [6]. Cela nous permet d'introduire comme critère le **potentiel de substitution**. Le service numérique évalué conduit-il de manière directe à une réduction ou à une augmentation des flux physiques ?

Troisièmement, les services numériques peuvent augmenter les flux physiques de manière directe, au travers des acquisitions de matériel, soit que le matériel courant devienne obsolète, soit que le service nécessite l'achat d'objets connectés. C'est le critère des **consommations induites**, soit le pendant des investissements induits côté fournisseur. Ce critère s'attellera à déterminer si le nouveau service contribue à l'obsolescence de terminaux existants (voir par exemple la polémique autour de l'incompatibilité à Windows 11 des machines précédemment compatibles Windows 10) ou à l'achat d'équipements complémentaires incluant des objets connectés ou des périphériques [24]. Or, les impacts liés aux objets connectés, dont le nombre croit rapidement, reste sous-exploré, en particulier dès lors qu'il concerne leur fabrication ou leur démantèlement [21].

Quatrièmement, les services numériques s'accompagnent souvent d'un effet rebond, en particulier dès lors que sa production est optimisée et en facilite la disponibilité, notamment sur un plan financier [20]. Cependant, le concepteur peut faire le choix de déployer des mécanismes incitatifs, soit pour encourager l'usage du service, soit pour en limiter l'utilisation au strict nécessaire [55]. Les interfaces peuvent ainsi encourager les usages compulsifs à l'aide de *dark patterns* ou de designs cornucopiens [5], ou au contraire pousser à la modération des usages à l'aide de *green nudges* [44, 57]. Ceux-ci peuvent passer par des indicateurs, soit conséquentiels (si l'impact est directement proportionnel à la consommation), soit attributifs (si l'impact est extrapolé depuis un calcul par unité fonctionnelle). Dans le second cas, la corrélation entre consommation et impact peut être faible. C'est notamment le cas des volumes de données où la consommation des équipements est essentiellement déterminée à leur mise sous tension. Cependant, même si l'indicateur est techniquement faux, il fournit une incitation utile à la tempérance des usages qui permet de limiter la pression sur les capacités des infrastructures. Ces dimensions sont couvertes par le critère de **tempérance des usages**.

3.2 Critères liés à la production

Premièrement, le déploiement d'un nouveau service numérique pose la question des **investissements induits**, soit le pendant des consommations induites côté utilisateur. Peut-il fonctionner sur des équipements pré-existants ou nécessite-t-il l'investissement dans de nouvelles capacités ? L'utilisation de grands modèles de langage s'accompagne ainsi d'investissements massifs en capacité de calcul destinées à l'entraînement des modèles et à l'inférence [28], ce qui occasionne des tensions d'approvisionnement [41].

Deuxièmement, le développement des services en ligne en général, et plus récemment des fournisseurs de *cloud computing* (Amazon Web Service, Google Cloud, Microsoft Azure…) a vu le développement des centres de données [55]. Se pose donc la question de l'**optimisation des infrastructures**. Cette optimisation passe par, d'une part, le matériel utilisé dans le centre de données, incluant le taux d'utilisation des machines et leurs performances [4], d'autre part, la conception du centre de données, incluant les dispositifs de refroidissement, son alimentation en énergie ou encore sa localisation [40, 42, 46].

Troisièmement, avoir des infrastructures efficaces n'est pleinement satisfaisant que si les logiciels qu'elles font tourner ont eux-mêmes été optimisés. Ce critère est donc relatif à l'**optimisation des logiciels**. Cette dernière passe par le choix de langage de programmation efficients [43], généralement compilés (C, C++, Rust, Go…) ou hybrides (Java, Scala...), la sélection d'algorithmes performants [14], mais aussi la mise en place de tests de performance [7] ou encore la limitation du périmètre fonctionnel du logiciel [5, 8].

Quatrièmement, la réduction du poids des infrastructures passe par un **traitement raisonné des données** [55]. Plus de données signifie davantage d'espace de stockage, davantage de capacités de calcul et des mises à jour des infrastructures réseaux. Limiter la taille des données traitées permet donc de travailler à équipements constants (terminaux, réseaux, centres de données), dès lors d'éloigner la nécessité de renouvellement des matériels ou d'accroissement des infrastructures.

Cinquièmement, les modèles d'intelligence artificielle deviennent un sujet d'inquiétude sur le plan environnemental. Schwartz et ses co-auteurs opposent ainsi ce qu'ils appellent la « Red AI », caractérisée par des coûts prohibitifs, un impact environnemental négatif et une élévation des barrières à l'entrée pour la recherche, et la « Green AI », plus inclusive et respectueuse de l'environnement [45]. Les poids environnementaux et financiers sont les deux versants d'une même pièce. En effet, les modèles, de plus en plus grands, nécessitent des capacités croissantes de calcul pour leur entraînement et leur utilisation, ce qui pousse ainsi à l'utilisation de modèles plus spécialisés et plus petits [56]. Les chercheurs appellent ainsi à moins se focaliser sur la précision et davantage sur l'efficacité des modèles. Cela permet d'introduire un dernier critère : la **frugalité du modèle d'IA**. Le service numérique s'appuie-t-il sur un modèle connexionniste de grande taille ou sur un modèle de *machine learning* plus classique ? L'utilisation de modèles au plus juste est-elle encouragée ?

4 Analyser la sobriété d'un portefeuille de services numériques

Les critères ci-dessus permettent de composer une grille synthétique où les deux branches (consommation, production) sont ensuite décomposées en critères évaluables sur une échelle de Likert [-2, +2] documentée. Associés à une échelle de notation (cf. « Annexe 1 »), ces critères, combinés à une pondération ajustable, permettent de construire une grille permettant d'obtenir un score global de sobriété numérique, associé à des scores partiels liés à la dimension « consommation » et à la dimension « production » du service numérique. L'exemple ci-dessous porte sur l'espace de stockage en ligne kDrive (cf. Tableau 2), avec une pondération des critères égale à 1.

Tableau 2. Évaluation de la sobriété numérique du service numérique kDrive (Infomaniak).

Critère	Score	Justification	
kDrive est un espace de stockage en ligne basé sur un *cloud* souverain suisse performant, prolongeant le matériel utilisateur, mais sans dispositif limitant automatiquement l'usage.			**B**
Consommation			**B**
Potentiel de régénération	0	Potentiel régénératif limité. Impact positif sur la collaboration, et amélioration de la souveraineté sur les données (base open-source, respect du RGPD et du nLPD).	
Potentiel de substitution	1	Réduction du besoin en disques externes individuels. Limitation des impressions et doublons. Moindre transport physique de documents.	
Consommations induites	1	Pas de nouveau matériel requis, support multi-OS. Besoin moindre en stockage local, ce qui peut prolonger la durée de vie des équipements existants.	
Tempérance des usages	0	Interface neutre ; aucun *dark pattern* ni fonctionnalité limitant automatiquement l'usage.	
Production			**A**
Investissements induits	0	Utilisation de centre de données existants ou mutualisation de plusieurs services, d'où un investissement supplémentaire marginal.	
Optimisation des infrastructures	2	Effort dans la conception des centres de données (cf. D4) : faible PUE, refroidissement par air extérieur, utilisation d'énergie renouvelable, réutilisation de la chaleur.	
Optimisation des logiciels	1	Code optimisé (éco-conception), compression automatique, profilage continu, dette technique maîtrisée selon documentation interne.	
Traitement raisonné des données	1	Volumes dictés par les utilisateurs. Conservation paramétrable. Pas de collecte systématique mais données potentiellement volumineuses.	
Frugalité du modèle IA	2	Aucun modèle IA gourmand. Aucune nécessité de processeurs spécialisés.	

Tableau 3. Quadrant magique de la sobriété numérique.

Consommation \ **Production**	Faible optimisation	Forte optimisation
Usage tempéré	Sobriété apparente Expérience utilisateur vertueuse masquant une empreinte infrastructurelle lourde.	Sobriété exemplaire Idéal-type caractérisé par la limitation des usages et la frugalité de l'infrastructure.
Usage compulsif	Voracité numérique Usages et infrastructures cumulativement problématiques avec un impact global défavorable.	Frugalité paradoxale Système technique optimisé mais usages compulsifs, soit une sobriété partielle, contrastée.

Avec un score de consommation égal à 0,5 (B) et un score de production égal à 1,5 (A), kDrive ressort comme un service numérique plutôt sobre. Les scores de consommation et de production, suivant qu'ils sont en moyenne inférieurs ou supérieurs à zéro, permettent de positionner chaque service numérique évalué dans un quadrant (cf. Tableau 3). Ce quadrant permet par exemple de visualiser l'état d'un portefeuille de services numériques au sein d'une organisation de manière à identifier les priorités d'amélioration. Cet outil reprend la logique similaire et complémentaire à celle adoptée par Sneed [48] pour planifier les opérations d'amélioration sur des systèmes hérités (priorisation fonction du caractère critique de l'application et de sa qualité, notamment structurelle).

La situation, d'une part, côté consommation, d'autre part, côté production, permet d'identifier quatre quadrants. Le plus problématique, soit la **voracité numérique**, se caractérise par des usages compulsifs et une lourde infrastructure. Le plus vertueux, soit la **sobriété exemplaire**, ressort comme un idéal-type de la sobriété numérique, marqué par la sobriété des usages et la frugalité des infrastructures utilisées. Deux cas intermédiaires sont présentés : la **sobriété apparente** et la **frugalité paradoxale**. Pour le premier, les usages sont limités mais l'infrastructure devrait être davantage optimisée. Dans le second, l'usage est compulsif mais est partiellement compensé par l'efficacité du système technique sous-jacent.

5 Cas particulier des services numériques basés sur l'IA générative

Étudier l'impact environnemental des IA génératives nécessite de choisir des indicateurs. Deux sont couramment utilisés : les émissions de CO_2 (bilan carbone) et la consommation d'eau. Combien tel ou tel modèle a-t-il provoqué d'émissions de CO_2 pour son entraînement ? La découpe en « scopes » (« scope 1 », « scope 2 », « scope 3 ») se retrouve tant pour le calcul du bilan carbone que pour l'estimation des impacts sur la ressource « eau » [9]. Ainsi, l'entraînement du modèle GPT-3 aurait nécessité 1,3 GWh, soit la consommation annuelle moyenne de 120 foyers étasuniens, occasionnant l'émission de 522 tonnes de CO_2 [50]. Quant à BLOOM, l'énergie consommée pour son entraînement est estimée à 433 MWh [39]. Concernant l'eau, Li et ses co-auteurs attribuent ainsi à une conversation avec ChatGPT une consommation de l'ordre de 50cl [35]. Plus précisément, ce chiffre de 50cl y représente la quantité d'eau nécessaire, dans le « scope 1 » et le « scope 2 », pour entraîner le modèle puis produire les réponses à un nombre de *prompts* compris entre 10 et 70 requêtes (cet intervalle de valeurs a légèrement évolué au gré des versions successives de la prépublication). Cette valeur peut être comparée au chiffre de 0,26 millilitres (soit environ 5 gouttes d'eau) revendiqué par Google pour son modèle Gemini [16]. Cette différence s'explique par trois facteurs. Premièrement, le calcul de la consommation ne couvre que l'inférence. Cela s'explique notamment par le fait que l'étude de Elsworth et ses co-auteurs [16] vise aussi, sur un plan plus marketing que scientifique, à mettre en évidence l'efficacité revendiquée des puces TPU conçues par Google comme alternative aux GPU de NVIDIA, le fournisseur dominant actuellement le marché

des puces dédiées à l'IA. Deuxièmement, leur étude se restreint au « scope 1 » pour le calcul de la consommation en eau (au contraire de l'étude de Li *et al.* [35], prenant en compte le « scope 1 » et le « scope 2 »). Troisièmement, alors que l'étude de Li et ses co-auteurs concerne GPT-3 [35], un modèle sorti en 2020, celle de Elsworth et ses co-auteurs [16] exploite les données internes relatives aux modèles les plus récents, sur lesquels des optimisations ont été réalisées pour en améliorer l'efficacité. Le secteur de l'IA générative est en effet marqué par les économies d'échelle [56]. Les coûts financiers liés à l'entraînement doivent être amortis sur une base d'utilisateurs la plus large possible (plus de 700 millions d'utilisateurs actifs pour ChatGPT [12]) car il s'agit d'un coût fixe. Les coûts financiers associés à l'inférence constituent par contre un coût variable : plus la base d'utilisateurs augmente, et plus le coût d'inférence l'emporte sur le coût d'entraînement. Google met donc en évidence l'innovation typique d'une phase transitionnelle visant à réduire le coût de production du service plutôt que d'en améliorer les performances pour l'utilisateur [55].

Ces deux exemples permettent par ailleurs d'interroger la provenance des données nécessaires à ce genre de comparaison. Les rapports environnementaux des *bigtechs* (p. ex. Google, Meta et Microsoft) sont riches en informations. Cependant, la transparence totale y est peu ou prou faite sur le coût environnemental des modèles. Les analyses d'impact s'appuient donc sur trois stratégies, non sans conséquence sur leur fiabilité. Premièrement, le calcul d'impact peut être réalisé par des chercheurs indépendants, suite à un lent travail de collecte de données et d'approximations, permettant d'élaborer un modèle de calcul de coûts, bien documenté mais nécessairement imprécis. Deuxièmement, le calcul d'impact peut être estimé par des équipes liées aux industriels (Google, Microsoft, OpenAI…). Cette configuration conduit à un accès privilégié à des données internes, ce qui autorise une analyse actualisée, plus proche de la réalité, mais aussi à de possibles conflits d'intérêt, par exemple pour présenter certaines informations sous un jour plus favorable. Troisièmement, le calcul d'impact peut être réalisé sur des modèles ouverts développés au sein de consortiums de recherche. La transparence est dès lors de mise. Cependant, ce type de modèle n'est pas soumis aux contraintes, notamment d'optimisation, d'un modèle réellement utilisé en production. L'analyse d'impact de GPT-3 [35, 36] correspond à la première configuration, celle de GLaM [42] ou de Gemini [16], à la seconde, et celle de BLOOM [39], à la troisième. L'évaluation chiffrée des impacts se révèle donc complexe du fait, d'une part, du manque de chiffres [37], d'autre part, des difficultés de comparaison.

Sur le plan de la consommation du service, le potentiel de substitution des IA génératives mériterait un examen attentif dès lors que l'automatisation de la production de textes ou d'images pourrait, dans certains contextes, conduire à des économies de ressources [52]. Le secteur de la production cinématographique, qui envisage déjà la production virtuelle comme moyen de réduire les transports, le temps de tournage et les besoins logistiques [31], pourrait par exemple bénéficier de modèles génératifs performants dédiés à la production de vidéos. Force est cependant de constater que les modèles génératifs font davantage naître des espoirs de gains de productivité, par exemple en programmation, même si ces derniers tardent à se concrétiser dans la pratique [10].

Sur le plan de la production du service, l'IAG appelle une attention particulière en matière d'investissements induits, d'optimisation des infrastructures et de frugalité des modèles IA. Premièrement, l'essor de l'IA générative entraîne de nombreux investissements pour l'acquisition de puces permettant l'entraînement et l'inférence [41]. Ainsi OpenAI est par exemple engagé dans un partenariat stratégique avec NVIDIA tandis qu'il a annoncé acheter pour 10 milliards de dollars de puces auprès de Cerebras. Deuxièmement, et même si l'optimisation des infrastructures fait partie des bonnes pratiques sectorielles, l'essor de l'IA connexionniste fait naître de nouveaux besoins en capacités de stockage et, surtout, de calcul, d'une part, pour l'entraînement des modèles, d'autre part, pour leur utilisation (inférence) au sein de systèmes techniques (agents conversationnels, moteurs de recherche ou applications spécialisées). L'IEA prévoit ainsi une croissance régulière de la consommation électrique des centres de données, autour de 15 % par an (ce qui aboutirait à un doublement de la consommation après 5 ans). Par ailleurs, le renforcement de capacités existantes conduit à de fortes disparités locales ou régionales [28]. Conséquence de leur concentration et de leur gigantisme, les installations les plus massives pèsent ainsi sur les infrastructures d'approvisionnement en électricité et en eau potable. Ainsi, la municipalité de West Des Moines, où Microsoft possède un centre de données sur lequel a été entraîné GPT-4, et ses services de distribution d'eau potable conditionnent l'approbation de tout futur centre de données à l'adoption de technologies réduisant drastiquement les pics de consommation d'eau par rapport aux projets actuels de Microsoft. Troisièmement, la réduction des impacts est possible à condition d'utiliser des modèles frugaux. Les grands modèles de langage (LLM, *Large Language Models*) coûtent cher, tant pour l'entraînement des modèles (p. ex. GPT) que lors de leur utilisation (inférence) au sein d'une application (p. ex. ChatGPT). L'optimisation des modèles peut donc porter sur l'entraînement (p. ex. nettoyage des jeux de données, réduction du nombre de paramètres et réduction du poids des paramètres) et l'inférence (p. ex. exécution partielle des modèles, utilisation de modèles plus petits pour des besoins plus simples et utilisation de puces spécialisées comme les LPU). Actuellement, trois types de modèles coexistent : les LLM (*Large Language Models*), les LRM (*Large Reasoning Model*) et les SLM (*Small Language Models*). Les LRM sont des LLM, généralement de grande taille, et optimisés pour le raisonnement. Ce sont les vaisseaux amiraux de modèles proposés. À l'opposé, les SLM sont des LLM plus légers, avec peu de paramètres, adaptés à la réalisation de tâches spécifiques. L'optimisation de l'impact environnemental des IA génératives passe ainsi par l'utilisation d'un modèle au plus juste (p. ex. routeur dans GPT-5) et la spécialisation de modèles. Les IAG se développent ainsi selon deux trajectoires [56]. La première

comporte des IA génératives généralistes (p. ex. ChatGPT) proposant souvent l'accès à deux types de modèles (un LLM et un LRM ; p. ex. GPT-4o et o1 ou DeepSeek-V3 et DeepSeek-R1 : cf. [15]), lourdes mais sensibles aux économies d'échelle (logique « *winner-takes-all* »). La seconde comporte des IA spécialisées, notamment basées sur des modèles ouverts (open-weight, open-source), plus rudimentaires mais sensiblement moins lourds. Par exemple, Perplexity a progressivement construit une expertise interne en matière d'hébergement de modèles ouverts (API interne « pplx-api ») puis de *finetuning* (modèles Sonar basé sur des modèles ouverts Mistral, Llama ou DeepSeek). Perplexity s'appuie sur l'infrastructure *cloud* d'Amazon Web Service ainsi que sur des prestataires spécialisés comme Cerebras, en plus de disposer d'un soutien technique de NVIDIA. Cet investissement s'est fait progressivement et a abouti à la famille de modèles Sonar, proposés comme LLMaaS mais aussi utilisés en interne. En particulier, l'utilisation de modèles orientés « réponses factuelles » a permis une réduction des coûts d'inférence sans perte de satisfaction des utilisateurs comparativement aux modèles propriétaires précédemment utilisés.

Tableau 4. Évaluation de la sobriété numérique du service numérique Euria (Infomaniak).

Critère	Score	Justification	
Euria est un agent conversationnel associé à la suite collaborative en ligne kDrive basé sur des LLM ouverts.			**C**
Consommation			**B**
Potentiel de régénération	-2	Potentiel régénératif absent. Amélioration de la productivité et substitution de services existants.	
Potentiel de substitution	-1	Optimisation de la recherche d'information mais nouvel usage numérique intensif.	
Consommations induites	0	Fonctionnement en ligne sur des terminaux existants.	
Tempérance des usages	0	Interface neutre ; aucun *dark pattern* ni fonctionnalité limitant automatiquement l'usage.	
Production			**C**
Investissements induits	-1	Utilisation de centre de données existants mais nécessité de capacités GPU supplémentaires.	
Optimisation des infrastructures	2	Effort dans la conception des centres de données (cf. D4).	
Optimisation des logiciels	1	Peu d'informations publiques sur l'éco-conception logicielle spécifique d'Euria. Bonnes pratiques standards.	
Traitement raisonné des données	-1	Minimisation de la collecte de données (p. ex. données à caractère personnel) mais pas d'absence de données massives.	
Frugalité du modèle IA	-1	Orchestration au plus juste de grands modèles de langage.	

Le Tableau 4 propose une évaluation de la sobriété numérique du service Euria, un agent conversationnel basé sur des modèles ouverts tels que Qwen, intégré à la suite collaborative kDrive (cf. Tableau 2).

6 Conclusion

La croissance des terminaux et la part « invisible » des impacts (réseaux, *datacenters*, fabrication et fin de vie) confirment que la digitalisation relève souvent d'un transfert de support plutôt que d'une dématérialisation. Dès lors, l'enjeu central n'est pas uniquement d'optimiser l'existant, mais d'abord de définir ce qui mérite d'être produit et consommé : quelles fonctions sont réellement utiles, quels usages doivent être encouragés, lesquels doivent être évités. Dans cette perspective, cette recherche propose une grille d'évaluation (par critère de notation) de la sobriété numérique mobilisable en amont de l'innovation (*fuzzy front end*), structurée en deux dimensions : consommation (caractère régénératif, substitution des flux physiques, consommations induites et tempérance des usages) et production (investissements induits, optimisation des infrastructures et des logiciels, minimisation des données et frugalité du modèle d'IA). Le quadrant associé facilite la comparaison entre services et la priorisation des actions, notamment face aux situations de voracité numérique. Les résultats soulignent que l'efficacité ne garantit pas la sobriété : sans tempérance, l'optimisation alimente l'effet rebond. L'IA générative cristallise ce risque : malgré des gains techniques, elle induit souvent des investissements en calcul et des pressions locales sur l'énergie et l'eau. D'où l'intérêt de privilégier le modèle au plus juste, des architectures permettant de réduire la taille des modèles (p. ex. RAG) et des infrastructures performantes, afin d'aligner l'innovation numérique avec les limites planétaires. L'outil proposé doit encore faire l'objet d'une validation du choix, de la couverture et de l'indépendance des critères. Sa mise à l'épreuve dans des cas d'usage concrets permettrait en outre de préciser les modalités d'application de la grille et de réduire la part de subjectivité dans l'évaluation.

7 Références

[1] Acquier, A., Mayer, J., & Valiorgue, B. (2024). Introduction. Anthropocène, limites planétaires et nouvelles frontières des sciences de gestion. Revue Française de Gestion, 315(2), 11-36. https://doi.org/10.1684/rfg.2024.18.

[2] Aggeri, F. (2023). Quelles innovations pour un monde de post-croissance ?. L'Économie politique 2023/2 N° 98, Éditions Alternatives économiques, pp. 42-57. https://shs.cairn.info/article/LECO_098_0042/pdf.

[3] Aurez, V. et Georgeault, L. (2016). Économie circulaire : système économique et finitude des ressources. De Boeck Supérieur. ISBN : 978-2-8073-0148-1.

[4] Barroso, L. A., & Hölzle, U. (2007). The case for energy-proportional computing. Computer, 40(12), 33-37. https://doi.org/10.1109/MC.2007.443.

[5] Beignon, A., Tabard, A., & Maudet, N. (2026, April). How Interface Design Choices Lead to Indirect Environmental Impacts Through Use Intensification. In CHI Conference on Human Factors in Computing Systems. https://dx.doi.org/10.1145/3772318.3790305.

[6] Berthoud, F., & Parry, M. (2010). Évaluation des impacts environnementaux de l'informatique. Quels outils ? Quelles limites ?. Terminal. Technologie de l'information, culture & société, (106-107). https://doi.org/10.4000/terminal.1794.

[7] Bondi, A. B. (2015). Foundations of software and system performance engineering: process, performance modeling, requirements, testing, scalability, and practice. Pearson Education. ISBN : 978-0-321-83382-2.

[8] Bordage, F. (2019). Sobriété numérique : Les clés pour agir. Buchet-Chastel. ISBN : 978-2283032152.

[9] Bouveret, S., Bugeau, A., Orgerie, A. C., & Quinton, S. (2024). De l'eau dans les nuages. Annales des Mines-Enjeux Numériques, 27(3), 41-48. https://doi.org/10.3917/ennu.027.0041.

[10] Calvino, F., Reijerink, J., & Samek, L. (2025). The effects of generative AI on productivity, innovation and entrepreneurship (OECD Artificial Intelligence Papers, No. 39). OECD Publishing. https://doi.org/10.1787/b21df222-en.

[11] Čegar, S., Drezgić, S., & Čišić, D. (2024). Exploring the Regenerative Economy: A Comprehensive Review of Literature. Journal of Regenerative Economics, 1(1), 1-40. https://crossreis.com/journal/JRE-Vol1-Issue1-2024.pdf.

[12] Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How people use chatgpt (No. w34255). National Bureau of Economic Research. https://doi.org/10.3386/w34255.

[13] Cisco (2021). VNI Complete Forecast Highlights. https://www.cisco.com/ (...) /Global_2021_Forecast_Highlights.pdf.

[14] Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2022). Introduction to algorithms. MIT press. ISBN : 978-0-262-03384-8.

[15] Deng, Z., Ma, W., Han, Q. L., Zhou, W., Zhu, X., Wen, S., & Xiang, Y. (2025). Exploring DeepSeek: A Survey on Advances, Applications, Challenges and Future Directions. IEEE/CAA Journal of Automatica Sinica, 12(5), 872-893. https://doi.org/10.1109/JAS.2025.125498.

[16] Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S., ... & Manyika, J. (2025). Measuring the environmental impact of delivering AI at Google Scale. arXiv preprint arXiv:2508.15734. https://doi.org/10.48550/arXiv.2508.15734.

[17] Faucheux, S., Hue, C., et Nicolaï, I. (2010). TIC et développement durable : Les conditions du succès. De Boeck. ISBN : 978-2804103712.

[18] FEMP (2024). Best Practices Guide for Energy-Efficient Data Center Design. Federal Energy Management Program, 26 juillet 2024. https://www.energy.gov/femp/articles/best-practices-guide-energy-efficient-data-center-design.

[19] Ferreboeuf, H., Efoui-Hess, M., Marrauld, L. et Lescop, C. (2020). Déployer la sobriété numérique. The Shift Project. https://theshiftproject.org/ (...) /2020/10/Deployer-la-sobriete-numerique_Resume_ShiftProject.pdf.

[20] Flipo, F., & Gossart, C. (2009). Infrastructure numérique et environnement. L'impossible domestication de l'effet rebond. Terminal. Technologie de l'Information, Culture & Société, (103-104). https://doi.org/10.4000/terminal.3093.

[21] Freitag, C., Berners-Lee, M., Widdicks, K., Knowles, B., Blair, G. S., & Friday, A. (2021). The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. Patterns, 2(9). https://doi.org/10.1016/j.patter.2021.100340.

[22] Google (2025). 2025 Environmental Report. Google, Juin 2025. https://sustainability.google/reports/google-2025-environmental-report/.

[23] Gossart, C. (2015). Rebound Effects and ICT: A Review of the Literature. In: Hilty, L., Aebischer, B. (eds) ICT Innovations for Sustainability. Advances in Intelligent Systems and Computing, vol 310. Springer, Cham. https://doi.org/10.1007/978-3-319-09228-7_26.

[24] Gould, P., Song, G., & Zhu, T. (2024). Environmental and Economic Impact of I/O Device Obsolescence. arXiv preprint arXiv:2412.20655. https://doi.org/10.48550/arXiv.2412.20655.

[25] Gregg, B. (2014). Systems performance: enterprise and the cloud. Pearson Education. ISBN : 978-0-13-339009-4.

[26] Guillard, V. (2021). Towards a society of sobriety: Conditions for a change in consumer behavior. Field Actions Science Reports. The Journal Of Field Actions, (Special Issue 23), 36-39. https://journals.openedition.org/factsreports/pdf/6590.

[27] Han, J., Li, B., Mukherjee, D., Chiang, C. H., Grange, A., Chen, C. et Bankoski, J. (2021). A technical overview of AV1. Proceedings of the IEEE, 109(9), 1435-1462. https://doi.org/10.1109/JPROC.2021.3058584.

[28] IEA (2025b), Energy and AI, International Energy Agency. https://www.iea.org/reports/energy-and-ai.

[29] IEA (2025a). Data Centres and Data Transmission Networks, International Energy Agency. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks.

[30] Kammoun, S., & Le Bas, C. (2020). L'innovation frugale peut-elle être disruptive?. Innovations, 63(3), 179-199. https://doi.org/10.3917/inno.063.0179.

[31] Khairulbahri, M. (2025). The application of the system archetypes in understanding GHG emissions: A case study in the US movie industry. Journal of Cleaner Production, 505, 145419. https://doi.org/10.1016/j.jclepro.2025.145419.

[32] Kotlarsky, J., Oshri, I., & Sekulic, N. (2023). Digital sustainability in information systems research: Conceptual foundations and future directions. Journal of the Association for Information Systems, 24(4), 936-952. http://doi.org/10.17705/1jais.00825.

[33] Kurp, P. (2008). Green computing. Communications of the ACM, 51(10), 11-13.

https://doi.org/10.1145/1400181.1400186.

[34] Le Bas, C. (2024). Les approches « low-tech » de l'innovation: une analyse critique des narratifs. Revue d'Économie Industrielle, 186(2), 175-197. https://doi.org/10.3917/rei.186.0175.

[35] Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making ai less' thirsty'. Communications of the ACM, 68(7), 54-61. https://doi.org/10.1145/3724499.

[36] Li, P., Yang, J., Islam, M. A., & Ren, S.(2023). Making AI less "thirsty": uncovering and addressing the secret water footprint of AI models. arXiv 2023. arXiv preprint arXiv:2304.03271. https://doi.org/10.48550/arXiv.2304.03271.

[37] Ligozat, A.-L., & Bugeau, A. (2025). Méthodes d'évaluation de l'empreinte de l'IA [Conference presentation]. GreenDays 2025, Rennes, France. HAL. https://hal.science/hal-05317063.

[38] Lombès, T., & Poubeau, B. (2014). Obsolescence programmée : mythes et réalité. Mines Paristech. http://www.annales.org/ (...) /Livre-Obsolescence- (...).pdf.

[39] Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2023). Estimating the carbon footprint of bloom, a 176B parameter language model. Journal of machine learning research, 24(253), 1-15. https://doi.org/10.48550/arXiv.2211.02001.

[40] Mytton, D. (2021). Data centre water consumption. npj Clean Water, 4(1), 11. https://doi.org/10.1038/s41545-021-00101-w.

[41] OECD (2025). Competition in artificial intelligence infrastructure (OECD Roundtables on Competition Policy Papers, No. 330). OECD Publishing. https://doi.org/10.1787/20758677.

[42] Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L. M., ... & Dean, J. (2022). The carbon footprint of machine learning training will plateau, then shrink. Computer, 55(7), 18-28. https://doi.org/10.1109/MC.2022.3148714.

[43] Pereira, R., Couto, M., Ribeiro, F., Rua, R., Cunha, J., Fernandes, J. P., & Saraiva, J. (2017). Energy efficiency across programming languages: how do energy, time, and memory relate?. In Proceedings of the 10th ACM SIGPLAN international conference on software language engineering (pp. 256-267). https://doi.org/10.1145/3136014.3136031.

[44] Schubert, C. (2017). Green nudges: Do they work? Are they ethical?. Ecological economics, 132, 329-342. https://doi.org/10.1016/j.ecolecon.2016.11.009.

[45] Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54-63. https://doi.org/10.1145/3381831.

[46] Sheme, E., Holmbacka, S., Lafond, S., Lučanin, D., & Frashëri, N. (2018). Feasibility of using renewable energy to supply data centers in 60° north latitude. Sustainable Computing: Informatics and Systems, 17, 96-106. https://doi.org/10.1016/j.suscom.2017.10.017.

[47] Sinkin, C., Wright, C. J., & Burnett, R. D. (2008). Eco-efficiency and firm value. Journal of accounting and public policy, 27(2), 167-176. https://doi.org/10.1016/j.jaccpubpol.2008.01.003.

[48] Sneed, H. M. (1995). Planning the reengineering of legacy systems. IEEE software, 12(1), 24-34. https://doi.org/10.1109/52.363168.

[49] Soares, I. V., Yarime, M., & Klemun, M. (2024). Balancing the trade-off between data center development and its environmental impacts: A comparative analysis of Data Center Policymaking in Singapore, Netherlands, Ireland, Germany, USA, and the UK. Environmental Science & Policy, 157, 103769. https://doi.org/10.1016/j.envsci.2024.103769.

[50] Sundberg, N. (2023). Tackling AI's Climate Change Problem. MIT Sloan Management Review, 65(2), 38-41. https://sloanreview.mit.edu/article/tackling-ais-climate-change-problem/.

[51] Szilas, N. (2024). Digital Sobriety: from Tips to Values. LIMITS workshop. https://computingwithinlimits.org/2024/papers/limits2024-szilar-digital-sobriety.pdf.

[52] Tomlinson, B., Black, R. W., Patterson, D. J., & Torrance, A. W. (2024). The carbon emissions of writing and illustrating are lower for AI than for humans. Scientific Reports (Sci Rep), 14(1), 3732. https://doi.org/10.17605/OSF.IO/YHTMQ.

[53] Vaija, S. et Philipot, É. (2020). L'importance des métaux rares pour le secteur des technologies de l'information et de la communication, le cas d'Orange. Annales des Mines-Responsabilité et environnement, 2, 24-28. https://doi.org/10.3917/re1.099.0024.

[54] Vanderbauwhede, W. (2023). Frugal Computing--On the need for low-carbon and sustainable computing and the path towards zero-carbon computing. arXiv preprint arXiv:2303.06642. https://doi.org/10.48550/arXiv.2303.06642.

[55] Viseur, R. (2025a). Sobriété numérique et traitement des données massives: vers des stratégies d'innovation plus durables. Terminal. Technologie de l'information, culture & société, (140), 61-83. https://doi.org/10.4000/13vhv.

[56] Viseur, R. (2025b). Des fonctionnalités à coût maîtrisé ? Le modèle A-U appliqué à l'IA générative. APIA (Applications Pratiques de l'Intelligence Artificielle), Dijon, France. https://hdl.handle.net/20.500.12907/52466.

[57] Viseur, R. (2026). Éthique de la gestion du consentement au traitement des données à caractère personnel : Comment les dark patterns permettent-ils d'orienter la prise de décision des internautes ? Revue ouverte d'ingénierie des systèmes d'information, Vol. 5. http://doi.org/10.21494/ISTE.OP.2025.1301.

[58] Yang, X., Li, Y., & Kang, L. (2020). Reconciling "doing good" and "doing well" in organizations' green IT initiatives: A multi-case analysis. International Journal of Information Management, 51, 102052. https://doi.org/10.1016/j.ijinfomgt.2019.102052.

Proposition d'un cadre argumentatif pour la Communication NonViolente

Safa Yahi[1]

[1]Université Aix-Marseille, LIS Lab

safa.yahi@lis-lab.fr

Résumé

La Communication Nonviolente (CNV en abrégé) est un processus de communication, développé par Marshal B. Rosenberg, qui vise à communiquer de telle sorte que les besoins de chacun soient entendus et respectés afin de trouver facilement une solution à une situation conflictuelle. Par ailleurs, l'argumentation est une approche d'Intelligence Artificielle qui peut être utilisée dans différentes situations telles que l'aide à la décision, la négociations et résolution de conflits. Le cadre argumentatif de référence est celui proposé par Phan Minh Dung. Il s'appuie sur un ensemble d'arguments et des relations d'attaques entre ces arguments. L'objectif est de trouver des ensembles d'arguments acceptables simultanément selon certaines sémantiques. L'objectif de cet article est de proposer un cadre ou un système argumentatif dédié à la CNV en adaptant celui de Dung.

Mots-clés

Argumentation, CNV

Abstract

NonViolent Communication (NVC) is a communication process developed by Marshall B. Rosenberg that aims to communicate in such a way that everyone's needs are identified, heard, and respected in order to find a solution to a conflictual situation.

Furthermore, argumentation is an approach in Artificial Intelligence that can be used in different situations such as conflict resolution, negotiation, and decision support.

The reference argumentation framework is the one proposed by Phan Minh Dung. It is based on a set of arguments and attack relations between these arguments. The objective is to identify sets of arguments that can be accepted simultaneously according to certain semantics.

The objective of this article is to propose an argumentation framework adapted to NVC, in which we define the notions of arguments, attacks, and extensions within the context of Nonviolent Communication.

Keywords

Argumentation, NVC

1 Introduction

La résolution des conflits constitue un enjeu important dans les interactions humaines, que ce soit dans un contexte professionnel ou personnel. Parmi les approches proposées afin de favoriser un dialogue constructif et de faciliter la résolution pacifique des désaccords, figure la Communication Nonviolente (CNV). Cette approche développée par Marshall B. Rosenberg propose un processus structuré visant à améliorer la qualité des échanges interpersonnels [1, 2]. La CNV peut être appliquée dans différents contextes : relations familiales (relation de couple, relation parent-enfant, etc) à l'école, au travail, pour des négociations diplomatiques, etc. Cette approche repose sur l'identification et l'expression claire des observations, des sentiments, des besoins et des demandes, dans le but de favoriser l'empathie, la compréhension mutuelle entre les interlocuteurs et faciliter ainsi la résolution de conflits. L'objectif central de la CNV est de permettre aux parties impliquées dans une situation conflictuelle d'identifier et de reconnaître les besoins de chacun afin de parvenir à des solutions respectueuses et satisfaisantes pour tous.

Parallèlement, en Intelligence Artificielle, les approches basées sur l'argumentation se sont largement développées. L'argumentation computationnelle offre des outils formels permettant de représenter et d'évaluer des arguments en présence de conflits d'opinion ou d'information. Ces approches sont utilisées dans divers domaines, notamment la résolution de conflits, la négociation, l'aide à la décision et les systèmes multi-agents [3].

Un cadre formel central dans ce domaine est le cadre argumentatif abstrait proposé par Phan Minh Dung [4]. Ce modèle représente un système argumentatif sous la forme d'un ensemble d'arguments reliés par des relations d'attaque. L'analyse de ces relations permet d'identifier des ensembles d'arguments acceptables simultanément selon différentes sémantiques d'acceptabilité, telles que les sémantiques admissibles, stables ou préférées. Ce cadre a servi de base à de nombreux travaux visant à modéliser des formes plus riches d'argumentation intégrant, par exemple, des préférences, des valeurs ou des structures internes d'arguments.

Dans ce contexte, il apparaît pertinent d'explorer les liens possibles entre les principes de la Communication NonViolente et les outils formels de l'argumentation. En effet, les

interactions décrites par la CNV peuvent être interprétées comme des échanges d'arguments fondés sur l'expression de besoins et de demandes, susceptibles d'entrer en conflit avec ceux d'autres interlocuteurs. Une modélisation argumentaire de ces interactions pourrait ainsi offrir un cadre formel permettant d'analyser ces conflits et de soutenir la recherche de solutions.

L'objectif de cet article est donc de proposer un cadre argumentatif adapté à la Communication NonViolente en adaptant celui de Dung en restant le plus fidèle possible au processus CNV décrit dans la littérature. Plus précisément, nous introduisons la notions d'arguments CNV-compatibles, nous définissons une relation d'attaque entre arguments CNV-compatibles. Nous introduisons une relation de reformulation d'un argument non CNV-compatible en un argument CNV-compatible. De plus, nous montrons l'utilité de considérer une nouvelle forme d'extension que nous appelons extension CNV-compatible.

Dans notre modèle, les conflits ne proviennent pas des besoins mais des stratégies utilisées pour les satisfaire. Une extension CNV-compatible correspond donc à un ensemble d'arguments dont les demandes sont compatibles et qui permettent de satisfaire les besoins de toutes les parties.

Le reste de cet article est structuré comme suit. Dans la section 2, nous présenterons les quatre composantes de la CNV. Nous rappellerons le cadre d'argumentation de Dung dans la section 3. Dans la section 4, nous proposerons une formalisation argumentative de la CNV. La section 5 conclura l'article.

2 Communication NonViolente

La Communication NonViolente (CNV) est une approche de communication développée par le psychologue américain Marshall B. Rosenberg dans les années 1960 [2, 1]. Elle vise à améliorer la qualité des relations humaines en favorisant l'écoute, l'empathie et l'expression authentique des besoins.

L'objectif principal de la CNV est de résoudre les conflits et communiquer sans agressivité ni jugement, tout en respectant les sentiments et les besoins de chacun.

Les composants principaux de la CNV sont :

- **Observation** : La première composante de la CNV consiste à donner une observation de la situation de manière objective et factuelle sans donner d'évaluation ou de jugement. Il importe d'éviter le jugement car ce dernier emmène potentiellement l'autre à résister à ce que nous disons et à se focaliser sur le fait de
  Par exmple, «Tu es toujours en retard.» exprime un jugement alors que «Tu es arrivé 20 minutes après l'heure prévue.» est un observation.
- **Sentiment** : La deuxième composante de la CNV revient à exprimer clairement nos sentiments et nos émotions suite à la situation décrite précédemment. Ceci nous permet d'établir plus facilement un lien avec les autres ce qui favorise la résolution de conflits.
- **Besoin** : La troisième composante de la CNV consiste à identifier les besoins non satisfaits ou non nourris qui ont généré les sentiments. Il existe une liste de besoins humains fondamentaux tels que la liberté, l'amour, la confiance, le soutient, etc. En CNV, on part du principe qu'exprimer ses besoins, au lieu de critiquer l'autre, permet d'établir une connexion avec l'autre et de résoudre plus facilement un conflit.
- **Demande** : La quatrième composante de la CNV nous invite à formuler des demandes contribuant à satisfaire nos besoins. Ces demandes doivent être claires, concrètes, réalisable et positives (nous demandons ce que nous voulons et non pas ce que nous ne voulons pas). Aussi, il importe de formuler une demande et non pas une exigence.

3 Argumentation

Le système d'argumentation de base selon [4] est défini sur un ensemble d'arguments et un ensemble d'attaques entre eux. Un argument fournit des raisons d'accepter ou de croire une affirmation, de choisir une option. Une attaque exprime des conflits entre les arguments. Chaque système d'argumentation peut être représenté par un graphe orienté où les nœuds représentent les arguments et les arêtes représentent les attaques entre eux.

Définition 1 (Dung's argumentation framework)**.** *Un système d'argumentation abstrait est une paire* AF = $\langle \mathcal{A}, \mathcal{R} \rangle$ *où* $\mathcal{A}$ *est un ensemble fini d'arguments et* $\mathcal{R}$ *est une relation binaire d'attaque définie sur* $\mathcal{A}\times\mathcal{A}$*. Étant donnés deux arguments* a *et* b*,* $a\ \mathcal{R}\ b$ *(ou bien* $(a, b \in \mathcal{R})$ *signifie que* a *attaque* b *(*a *est un attaquant de* b*).*

La sortie d'un AF sont des ensembles d'arguments acceptables, appelés extensions. Il existe plusieurs catégories d'extensions : admissibles, préférées, stables, etc. Elles sont appelées également sémantiques d'acceptabilité.

Étant donné un cadre argumentatif AF, il est possible qu'un argument donné de AF soit acceptable selon une sémantique donnée mais non acceptable selon une autre sémantique.

Définition 2. *Soit* $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ *un système d'argumentation et* $\mathcal{S}$ *un sous-ensemble d'arguments (càd* $\mathcal{S} \subseteq \mathcal{A}$*),*

- $\mathcal{S}$ *est sans conflit de* AF *ssi il n'existe pas d'arguments* $a, b \in \mathcal{S}$ *tels que* $a\ \mathcal{R}\ b$*.*
- $a \in \mathcal{A}$ *est acceptable par rapport à* $\mathcal{S}$ *ssi* $\forall b \in \mathcal{A}$ *tel que* $b\ \mathcal{R}\ a$*,* $\exists\ c \in \mathcal{S}$ *tel que* $c\ \mathcal{R}\ b$*.*
- $\mathcal{S}$ *est un ensemble admissible ssi il est sans conflits et chaque argument dans* $\mathcal{S}$ *est acceptable par rapport* $\mathcal{S}$*.*
- $\mathcal{S}$ *est une extension préférée ssi elle maximale par rapport à l'inclusion ensembliste.*
- $\mathcal{S}$ *est une extension stable ssi elle est sans conflit et attaque tout argument dans* $\mathcal{A}\backslash\mathcal{S}$*.*

Une fois les extensions définies, voici le statut d'un argument :

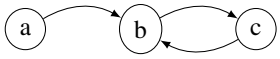


FIGURE 1 – Exemple d'un AF

Définition 3 (Status of arguments). *Soit* AF = $\langle \mathcal{A}, \mathcal{R} \rangle$ *un cadre d'argumentation et soient* $\varepsilon_1, \ldots, \varepsilon_n$ *ses extensions selon une certaine sémantique. Soit* $a \in \mathcal{A}$.

- a *est sceptiquement accepté ssi* $\forall i \in \{1 \ldots n\}, a \in \varepsilon_i$.
- a *est crédulement accepté ssi* $\exists i \in \{1 \ldots n\}$ *tel que* $a \in \varepsilon_i$.
- a *est rejeté* $\forall i \in \{1 \ldots n\}, a \notin \varepsilon_i$.

Exemple 1. *Considérons le cadre argumentatif* AF *donné par la Figure 1 où* $\mathcal{A} = \{a, b, c\}$, $\mathcal{R} = \{(a, b), (b, c), (c, b)\}$.
AF *possède les ensembles admissibles suivants :* $\emptyset$, *a, c, a, c.*
L'extension préférée de AF *est a, c, qui est également stable. Les arguments* a *et* c *sont acceptés de manière sceptique selon la sémantique préférée, tandis que l'argument* b *est rejeté.*

4 Communication NonViolente et Argumentation

Nous commençons par définir la notion d'un argument compatible avec le principe de la Communication NonViolente comme suit :

Définition 4. *Soit* O *un ensemble d'observations,* S *un ensemble de sentiments,* B *un ensemble de besoins et* D *un ensembles de demandes. Un argument CNV-compatible est un quadruple* $(o, s, b, d) \in O \times S \times B \times D$.

Intuitivement, la demande d est justifiée par le besoin b dont découle le sentiment s né suite à l'observation o.

En résolution de conflits en CNV, le premier objectif est que chacun exprime ses sentiments et ses besoins pour créer une connexion entre les personnes du conflit. Une fois la connexion établie, on cherche les stratégies qui satisfont les besoins de tout le monde. Chacun est invité à chercher, avec empathie, les besoins de l'autre, quelque soit la manière dont c'est exprimé même s'il s'agit d'un jugement. On cherche à reformuler les propos qui ne sont pas compatibles avec la CNV pour déterminer les sentiments et les besoins qu'ils expriment. Cette reformulation empathique est faite soit par une partie du conflit soit par un médiateur formel ou informel. Ainsi, tous les arguments sont importants même ceux qui ne sont pas CNV-compatibles, il convient de les écouter, de les analyser et de les reformuler en arguments CNV-compatibles. Ceci nous amène à considérer une nouvelle relation dans notre système d'argumentation : il s'agit de la reformulation empathique.

Intéressons nous à présent à la relation d'attaques entre arguments CNV-compatibles. En CNV, on suppose que les besoins humains sont universels et ne peuvent pas être incompatibles entre eux. De plus, on suppose que les observations sont fiables. D'ailleurs, on dit que l'on peut les vérifier avec une caméra. Les sentiments sont subjectifs et ne nous semblent pas attaquables. Par conséquent, nous définissons la relation d'attaques entre arguments CNV-compatibles comme une relation d'incompatibilité entre les demandes de ces arguments.

Définition 5. *Un argument* (o_1, s_1, b_1, d_1) *attaque un autre argument* (o_2, s_2, b_2, d_2) *si et seulement si les demandes* d_1 *t* d_2 *sont incompatibles càd irréalisables en même temps.*

Notons que cette relation d'attaque est, par définition, symétrique.
Comme les arguments non CNV-compatible sont sensés être reformulés et substitués par des arguments CNV-compatible, nous nous intéressons pas à des attaques impliquant des arguments non CNV-compatibles.
Voici à présent la définition d'un système d'argumentation en CNV.

Définition 6. *Un système d'argumentation en CNV est défini par le triplet* $< A, E, R >$ *où :*

- A *est un ensemble d'arguments qui peuvent être CNV-compatibles ou non CNV-compatibles.*
- E *est une relation de reformulation empathique d'un argument non CNV-compatible par un argument CNV-compatible.*
- R *est une relation d'attaques entre arguments CNV-compatibles (caractérisée par une incompatibilité de leurs demandes).*

Notons que nous ne nous intéressons pas aux relations d'attaques qui pourraient impliquer des arguments non CNV-compatibles. En revanche, nous nous focalisons sur leurs traductions CNV-compatibles qui reflèteraient autrement ces attaques.

Nous trouvons que calculer les extensions classiques (préférées, stables, etc) sur ce système d'argumentation ne nous permet pas de distinguer les arguments qui constituent une solution du conflit d'autres arguments qui le font pas. Ceci est démontré par l'exemple 2.
Par conséquent, nous proposons une nouvelle définition d'extension plus adaptée dans ce cadre comme suit :

Définition 7. *Soit* $< A, E, I >$ *un système d'argumentation CNV. Soit* S *un sous-ensemble d'arguments de* A *qui sont CNV-compatibles.* S *de une extension CNV-compatible ssi il vérifie les conditions suivantes :*

- *Il n'existe aucun argument non CNV-compatible dans* A *qui n'a pas été reformulé en un argument CNV-compatible.*
- S *est sans conflits (au sens classique de l'argumentation de Dung).*
- *Pour tout besoin d'un argument CNV-compatible dans* A, *il existe un argument dans* S *dont la demande le satisfait.*

Illustrons nos propos cet exemple.

Exemple 2. *Considérons une situation où des membres d'une famille échangent sur l'utilisation du téléphone par leur fils*

- *La maman dit à son fils : "Tu es toujours sur ton téléphone, arrête immédiatement". (**Argument** a_1)*
- *Le papa, adoptant une approche CNV, reformule le message de la mère comme suit : " Tu utilises ton téléphone depuis 2 heures. Nous sommes inquiets. Nous avons besoin d'être rassurés pour ta santé. Est-ce que tu peux arrêter l'utilisation de ton téléphone et aller faire travailler ton instrument de musique ?». (**Argument** a_2)*
- *La maman confirme le propos du papa.*
- *Le fils répond en CNV : « Quand je discute avec mon ami, je me sens épanouis. J'ai besoin de connexion sociale. Est-ce que je peux continuer à lui parler au téléphone ? (**Argument** a_3)*
- *A présent, les besoins de tout le monde sont identifiés. Ils peuvent plus facilement trouver des stratégies qui nourrissent tous ces besoins.*
- *Ainsi, la maman propose : "Que pense-tu d'inviter ton ami pour sortir ensemble au parc ? Ceci satisfait ton besoin de connexion sociale et notre besoin d'être rassurés pour ta santé. (**Argument** a_4)*
- *Le fils a approuve finalement cette stratégie et le conflit est résolu.*

Dans l'approche que nous proposons, l'argument a_1 n'est pas CNV-compatible. En effet, il comporte un jugement et non pas une observation. Il comporte aussi une exigence au lieu d'une demande. Il a été reformulé par un argument CNV-compatible, à savoir a_2 qui permet d'établir une meilleure connexion avec le fils. a_1 sera ainsi ignoré dans le calcul des extensions.

Par ailleurs, selon la définition 5, les arguments a_2 et a_3 s'attaquent mutuellement car les demandes correspondantes sont incompatibles et ne peuvent pas se réaliser en même temps. De la même manière, les arguments a_3 et a_4 sont incompatibles, ainsi que les arguments a_3 et a_4.
Dans ce cas, les extensions préférées selon Dung sont $\{a_2\}$, $\{a_3\}$ et $\{a_4\}$. Ce sont aussi des extensions stables.
Selon notre définition 7, a_4 est la seule extension CNV-compatible. Il s'agit bien d'une solution au conflit ce qui distingue cet argument des autres.

5 Conclusion et perspectives

Dans cet article, nous avons proposé un cadre argumentatif adapté à la Communication NonViolente (CNV) en nous appuyant sur les fondements de l'argumentation abstraite introduits par Phan Minh Dung. Cette formalisation permet de représenter de manière structurée les situations de conflit et d'identifier des ensembles d'arguments compatibles qui correspondent à des solutions respectant les besoins exprimés par les différentes parties.
Plusieurs perspectives de recherche peuvent être envisagées. Par exemple, nous pensons qu'il serait pertinent d'étudier le lien entre la CNV et l'argumentation basée sur les valeurs [5], l'argumentation bipolaire [6] ou l'argumentation basée sur le "ranking" [3]. Le lien avec le problème de satisfaction de contraintes (CSP) nous semble intéressant également à analyser.
Une autre perspective qui nous trouvons prometteuse concerne l'utilisation des modèles de langage de grande taille (LLM) pour identifier automatiquement les composantes de la CNV (observations, sentiments, besoins et demandes) à partir de dialogues naturels. De tels outils pourraient faciliter la reformulation d'arguments non compatibles avec la CNV et assister les processus de médiation et de résolution de conflits.

Références

[1] Marshall B. Rosenberg. *Les mots sont des fenêtres (ou bien ce sont des murs) : Introduction à la Communication NonViolente*. La Découverte, Paris, 2005. Traduction française de : Nonviolent Communication : A Language of Life.

[2] Marshall B. Rosenberg. *Nonviolent Communication : A Language of Life*. PuddleDancer Press, Encinitas, CA, 2003.

[3] Leila Amgoud and Jonathan Ben-Naim. Argumentation-based ranking logics. In Gerhard Weiss, Pinar Yolum, Rafael H. Bordini, and Edith Elkind, editors, *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015*, pages 1511–1519. ACM, 2015.

[4] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2) :321–358, 1995.

[5] Trevor J. M. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3) :429–448, 2003.

[6] Leila Amgoud, Claudette Cayrol, Marie-Christine Lagasquie-Schiex, and P. Livet. On bipolarity in argumentation frameworks. *Int. J. Intell. Syst.*, 23(10) :1062–1093, 2008.

SESSION 8 : Apprentissage et classification présidée parAstrid Klipfelr

Génération de Données Synthétiques Équitables et Préservant la Vie Privée via Autoencodeur Variationnel Basé sur le Clustering et Réseaux Antagonistes Génératifs

Malek Adouani[1], Zaineb Chelly Dagdia[2,3]

[1] Université Paris-Saclay, UVSQ, DAVID
[2] Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL
[3] Univ. Lille, Inserm, CHU Lille, U1286 – INFINITE – Institute for Translational Research in Inflammation

26 février 2026

Résumé

Le recours à l'apprentissage automatique en santé soulève des enjeux majeurs de biais et de confidentialité. Les modèles génératifs équitables reposent souvent sur des données étiquetées, limitant leur usage en non supervisé, tandis que les approches basées sur la confidentialité différentielle protègent les données sans éliminer nécessairement les biais latents. Les méthodes existantes n'optimisent pas conjointement équité et confidentialité sans supervision explicite. Nous proposons un cadre génératif hybride combinant un VAE avec clustering sous zCDP et un WGAN-GP. Le clustering dans l'espace latent structure les représentations et guide la génération en contexte non supervisé, tandis qu'un Fairness Critic pénalise les corrélations avec les attributs sensibles qui sont des attributs démographiques (par exemple, sexe, origine, âge). Les résultats montrent une réduction efficace des biais, une forte garantie de confidentialité et une bonne utilité des données synthétiques. Cet article correspond à résumé d'un travail publié à ECML PKDD'2025 [1]. L'article original est accessible à l'adresse suivante : https://hal.science/hal-05113907v1/document.

Mots-clés

Réseaux antagonistes génératifs, Apprentissage non supervisé, Atténuation des biais, Préservation de la vie privée.

1 Problématique et état de l'art

L'adoption croissante des modèles d'apprentissage automatique dans des domaines sensibles tels que la santé et la finance soulève des préoccupations majeures liées aux biais algorithmiques et à la confidentialité des données. Les modèles entraînés sur des données biaisées peuvent reproduire, voire amplifier, des inégalités existantes, tandis que la protection des informations individuelles, souvent formalisée par la confidentialité différentielle (DP) constitue une contrainte incontournable. Pourtant, concilier ces deux exigences demeure particulièrement complexe, notamment dans des contextes non supervisés où les données ne sont pas annotées. La littérature sur la génération de données synthétiques s'est principalement structurée autour de deux axes distincts. D'une part, les approches orientées équité, telles que FairGAN et TabFairGAN [4], introduisent des contraintes de parité statistique entre attributs sensibles (par exemple, sexe, origine, âge) afin de limiter les discriminations. Toutefois, ces méthodes reposent sur des annotations explicites pour guider le débiaisage, ce qui restreint leur applicabilité en absence d'étiquettes. D'autre part, les modèles axés sur la confidentialité, comme les DPGANs [3], intègrent des mécanismes de DP via l'injection de bruit dans les gradients ou les données de sorties du modèle. S'ils offrent des garanties formelles de protection de la vie privée, ils ne traitent pas explicitement les biais latents, susceptibles d'être reproduits dans les données synthétiques. Ainsi, les travaux existants répondent généralement à l'un ou l'autre de ces objectifs, mais rarement aux deux simultanément, révélant une lacune dans les cadres non supervisés. Pour y remédier, nous proposons Clust-VAE-WGAN-GP, un modèle génératif hybride combinant une structuration équitable de l'espace latent et des mécanismes formels de confidentialité, afin de produire des données synthétiques à la fois équitables et privées.

2 Méthode proposée

Notre modèle, Clust-VAE-WGAN-GP, est une architecture hybride en deux étapes.

Étape 1 : Autoencodeur Variationnel (VAE) Basé sur le Clustering avec zCDP : La première étape structure les données d'entrée et garantit la confidentialité ou un VAE encode les données dans un espace latent continu. Directement dans cet espace latent, nous appliquons K-Means pour regrouper les représentations similaires et capturer la structure inhérente des données sans supervision. Les étiquettes de clusters qui en résultent servent de signal conditionnel pour la deuxième étape. Pour la confidentialité, nous intégrons la zero-Concentrated Differential Privacy (zCDP) [2], une variante mathématiquement rigoureuse de la DP, en ajoutant un bruit gaussien calibré aux gradients lors de l'entraînement du VAE. Cette étape produit des représentations latentes structurées et privées.

Étape 2 : WGAN-GP avec Débiaisage Adversarial La seconde étape génère les données synthétiques finales. Nous utilisons un réseau antagoniste génératif (GAN) de Wasserstein avec Pénalité de Gradient (WGAN-GP) pour sa stabilité d'entraînement. Le générateur est conditionné par les étiquettes de clusters produites par le VAE, garantissant que les données générées s'alignent sur la structure découverte. L'innovation majeure ici est l'ajout d'un discriminateur d'équité (Fairness Critic, FC), un second discriminateur entraîné spécifiquement pour pénaliser les corrélations entre les données générées et des attributs sensibles prédéfinis. Le générateur est alors entraîné de manière adversariale non seulement pour produire des données réalistes, mais aussi pour produire des données indépendantes des attributs sensibles. Ce mécanisme de débiaisage adversarial impose l'équité de manière dynamique en absences d'annotations.

3 Expérimentions et Résultats

Nous avons évalué notre modèle sur six jeux de données du domaine de la santé : HIV, Stress, Obesity, Cardio, Diabetes et Pediatric, et pour chaque base on a fixé les attributs sensibles qui sont genre, origine, et âge. Les performances ont été comparées à des méthodes de référence en matière d'équité (FairGAN, TabFairGAN) et de confidentialité (DPGAN, RDP-CGAN). L'évaluation a porté sur : le réalisme des données, l'équité, et la confidentialité.

3.1 Réalisme et Utilité des Données

Sans contraintes de confidentialité, Clust-VAE-WGAN-GP est compétitif par rapport aux méthodes de comparaisons, il obtient les meilleures performances dans quatre des six métriques de réalisme pour les jeux de données Heart et Pediatric. Par exemple, sur la base Pediatric, notre méthode atteint les plus faibles valeurs de Discrépance moyenne maximale (MMD) (0.0010) et de distance de Wasserstein (0.238), et les scores les plus élevés de Score de probabilité par dimension (DWP) (0.5977), α-précision (0.741) et β-rappel (0.712). Cette performance est attribuée à l'inférence variationnelle du VAE et au mécanisme de conditionnement par clustering qui guide la génération.

3.2 Équité et Réduction des Biais

L'évaluation de l'équité a montré l'efficacité de notre approche pour atténuer les biais de manière non supervisée. Sur le jeu de données Pediatric, notre modèle a obtenu les écarts d'information mutuelle (MI) les plus faibles entre les clusters et les attributs sensibles (MI-Genre = 0.0010, MI-Âge = 0.0006), indiquant une indépendance quasi-totale. De plus, il a atteint le meilleur Score de Silhouette (0.5297) et le plus faible Indice de Davies-Bouldin (0.5025), confirmant une structure de clusters cohérente et non biaisée. Ces résultats valident l'apport du FC pour pénaliser activement les dépendances statistiques indésirables.

3.3 Préservation de la Confidentialité

Le modèle offre de solides garanties de confidentialité, avec un risque d'identifiabilité (EIR) très faible (entre 0.0100 et 0.0221 sur tous les jeux de données). Nous avons également analysé le compromis entre l'équité et la confidentialité. Les résultats montrent que lorsque le budget de confidentialité augmente (moins de bruit, donc moins de protection), les métriques d'équité comme le Score de Silhouette s'améliorent, indiquant que le débiaisage est plus efficace lorsque le signal des données n'est pas excessivement masqué par le bruit de la DP.

4 Conclusion

Nous avons présenté Clust-VAE-WGAN-GP, une nouvelle approche générative hybride qui impose simultanément l'équité et la confidentialité pour la génération de données synthétiques, en absence de label. En combinant un VAE basé sur le clustering avec zCDP et un WGAN-GP doté d'un débiaisage adversarial, notre approche parvient à générer des données de haute fidélité, équitables et privées. Les expérimentations sur six jeux de données démontrent sa supériorité par rapport aux modèles existants. Ce travail ouvre la voie à une génération de données plus éthique et fiable dans les domaines sensibles. Comme perspectives, nous envisageons d'explorer des techniques de clustering adaptatif pour améliorer davantage l'atténuation des biais.

Remerciements

Ce travail est financé et soutenu par : le programme Horizon Europe de l'Union européenne (MSCA, convention n° 101236749), les financements France 2030 RHU RECORDS (ANR-18-RHUS-0004), IHU SEPSIS (ANR-23-IAHU-0004), iRECORDS financé par ERA PerMed (JTC_2021), le Programme d'Investissements d'Avenir (I-SITE ULNE / ANR-16-IDEX-0004 ULNE), géré par l'Agence Nationale de la Recherche (n° I-KUL-22-005-ARCHIE-INFINITE), ainsi que d'un financement de l'Inserm et du Ministère de la Santé via l'appel MESSIDORE 2023 opéré par l'IReSP (AAP-2023-MSDR-341423).

Références

[1] Malek Adouani and Zaineb Chelly Dagdia. Fair and Privacy-preserving Synthetic Data Generation via Clustering-based Variational Autoencoder and Adversarially Debiased Wasserstein Generative Adversarial Networks with Gradient Penalty. In *ECMLPKDD 2025*, September 2025.

[2] M. Bun and T. Steinke. Concentrated differential privacy : Simplifications, extensions, and lower bounds. *arXiv preprint*, arXiv :1605.02065, 2016.

[3] G. O. Ghosheh, J. Li, and T. Zhu. A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Computing Surveys*, 55(8), 2023.

[4] A. Rajabi and O. O. Garibay. Tabfairgan : Fair tabular data generation with generative adversarial networks. *arXiv preprint*, arXiv :2109.00666, 2021.

Étude comparative de méthodes de vision par ordinateur pour la détection de quasi-doublons d'images de coraux

Alissa Chahine[1,2], Juliette Van Poulle[1,2]

[1] Université de Technologie de Compiègne, Département Génie Informatique
[2] Aubay, Lab'Innov

Résumé

La qualité du jeu de données est une préoccupation récurrente en vision par ordinateur. Cette étude évalue et compare différentes méthodes, classiques et profondes, de détection de quasi-doublons afin d'améliorer la qualité d'un jeu de données d'images de coraux. Les quasi-doublons sont définis comme des images distinctes présentant de faibles variations et représentant un même individu de corail. Leur présence dans un jeu de données réduit l'efficacité de l'apprentissage des modèles, peut biaiser l'évaluation de leur performance et augmente inutilement les besoins en stockage et en calcul. De plus, les images sous-marines présentent des altérations importantes. Ce facteur, combiné aux motifs spécifiques des coraux, confère à cette problématique un besoin important de robustesse et d'adaptation des méthodes de détection existantes.

Mots-clés

quasi-doublons, vision par ordinateur, hachage perceptuel, NetVLAD, ORB, DELF

Abstract

Dataset quality is a recurring concern in computer vision. This study evaluates and compares different near-duplicate detection methods, both classical and deep learning-based, to improve the quality of a coral image dataset. Near-duplicates are defined as distinct images exhibiting minor variations while representing the same coral individual. Their presence in a dataset reduces the effectiveness of model training, can bias performance evaluation, and unnecessarily increases storage and computational requirements. In addition, underwater images often present significant alterations. This factor, combined with the specific visual patterns of corals, highlights the need for robustness and adaptation of existing detection methods to address this challenge.

Keywords

near-duplicates, computer vision, perceptual hashing, NetVLAD, ORB, DELF

1 Introduction

Les quasi-doublons sont des images représentant le même contenu mais avec de légères variations, telles que des changements de rotation, d'échelle, de recadrage, de luminosité, d'équilibre des couleurs ou de point de vue. Elles ne sont pas identiques pixel par pixel, mais elles représentent visuellement la même scène ou le même objet.

La présence de ces images au sein d'un jeu de données peut poser problème, selon le contexte dans lequel elle se manifeste. Il faut donc distinguer deux situations fondamentalement différentes. D'une part, l'augmentation de données est une pratique standard et volontaire en vision par ordinateur : elle consiste à introduire délibérément des variantes transformées d'images existantes afin d'améliorer la capacité de généralisation des modèles. D'autre part, la redondance non maîtrisée au sein d'un corpus, c'est-à-dire la présence de quasi-doublons non identifiés, répond à une tout autre logique et peut, elle, introduire des biais indésirables. Lorsque des quasi-doublons apparaissent de manière non contrôlée à la fois dans les ensembles d'entraînement et de test, la performance du modèle peut être artificiellement gonflée : au lieu de généraliser à de nouvelles observations, le modèle retrouve une image quasi identique à celle vue lors de l'entraînement. Cela conduit à des métriques de précision trompeuses et affaiblit la robustesse réelle du système. Par ailleurs, une telle redondance peut causer une surreprésentation non souhaitée de certaines classes et entraîner un besoin inutile de stockage et de calcul. Cette problématique est critique dans le contexte de jeux de données de surveillance environnementale, où les images sont souvent acquises en séquence dans des conditions similaires, favorisant l'émergence naturelle de quasi-doublons. Cette étude a été réalisée suite à la demande de la Fondation Science4Reefs, qui œuvre à la préservation des récifs coralliens en Polynésie française, auprès de la société Aubay. Son objectif est d'évaluer la faisabilité de la détection automatique de telles images au sein du jeu de données d'images de coraux fourni par la Fondation Science4Reefs. L'approche proposée adapte et compare trois catégories de méthodes : le hachage perceptuel, les descripteurs globaux d'images (NetVLAD), et les descripteurs locaux reposant sur la détection de points-clés (ORB et DELF). La combinaison de ces approches offre une analyse comparative des compromis entre précision, robustesse aux transformations et coût computationnel. Cette étude se concentre ainsi sur la capacité de chaque méthode à retrouver des quasi-doublons sous des transformations réalistes.

2 Contexte

Cette section pose les bases de l'étude comparative. On y présente d'abord les travaux existants dans la littérature, en détaillant quatre familles de méthodes d'extraction de descripteurs d'images retenues : le hachage perceptuel, NetVLAD, ORB et DELF. On décrit ensuite le jeu de données fourni par la Fondation Science4Reefs, composé de photographies sous-marines de coraux de Moorea, puis les métriques d'évaluation (Recall@K, mAP, Recall et Pureté) utilisées pour mesurer les performances des approches étudiée sont définies, avant de présenter la méthodologie mise en place pour l'étude comparative.

2.1 Travaux de la littérature

La détection de quasi-doublons est particulièrement étudiée en vision par ordinateur, bien que peu documentée pour les images aquatiques. Elle repose sur l'extraction de descripteurs d'images, étape centrale, et le calcul de leur similarité pour identifier les correspondances. Ainsi, cette section présente quatre méthodes d'extraction de descripteurs d'images retenues pour cette étude comparative : le hachage perceptuel, les descripteurs globaux et les descripteurs locaux. Leur sélection s'appuie un sur travail exploratoire de recensement et d'analyse critique de publications concernant la détection de quasi-doublons et la reconnaissance visuelle de lieux. Un benchmark a permis l'analyse comparative de ces méthodes à l'aide de plusieurs critères, notamment l'adéquation avec les ressources disponibles et leur robustesse face aux variations attendues.

2.1.1 Hachage perceptuel

Le hachage perceptuel est une technique qui consiste à transformer une image en une empreinte numérique de taille fixe, appelée hash. Cette empreinte résume les caractéristiques visuelles principales de l'image et permet de comparer efficacement des images entre elles sans avoir à les comparer pixel par pixel. Contrairement aux fonctions de hachage cryptographique classiques, qui produisent des empreintes totalement différentes dès qu'un seul pixel est modifié, le hachage perceptuel a pour propriété de préserver la similarité visuelle : deux images proches visuellement produisent des hashes proches, mesurables par la distance de Hamming. Plusieurs variantes existent dans la littérature, parmi lesquelles aHash (average Hash), dHash (difference Hash) et pHash (perceptual Hash) [8]. Ces méthodes se distinguent par la manière dont elles résument le contenu de l'image, respectivement par la moyenne des niveaux de gris, les gradients horizontaux locaux, et les coefficients de basse fréquence d'une transformée en cosinus discrète (DCT). Leur principal avantage est leur faible coût computationnel, qui les rend adaptées au traitement de jeux de données à grande échelle. En revanche, leur robustesse reste limitée face à des transformations géométriques importantes telles que les rotations ou les recadrages agressifs.

2.1.2 NetVLAD

NetVLAD [1] est une méthode d'apprentissage profond d'extraction de caractéristiques globales d'images. Elle est basée sur un réseau de neurones convolutifs (CNN) intégrant une couche d'agrégation de type Vector of Locally Aggregated Descriptors (VLAD) qui permet un apprentissage de bout en bout. Bien que le terme « NetVLAD » corresponde à la couche d'agrégation, il désigne, dans cette étude, la méthode complète. Le CNN est tronqué au niveau de sa dernière couche convolutive et sert d'extracteur de descripteurs locaux, robustes aux variations d'éclairage et d'angle tout en conservant leurs caractéristiques spatiales. La couche NetVLAD permet d'agréger les descripteurs locaux extraits par le CNN en une représentation globale et compacte de l'image. Il s'agit d'une généralisation de la couche VLAD classique, qui y introduit un mécanisme d'attribution plus souple. Les descripteurs d'images globaux produits sont adaptés à une reconnaissance visuelle de lieux sous faible supervision et restent robustes face à des transformations courantes dans les milieux aquatiques, ce qui rend la méthode fortement pertinente pour cette étude.

2.1.3 ORB

ORB (Oriented FAST and Rotated BRIEF) [6] est un descripteur local rapide et robuste, conçu comme une alternative libre de droits à SIFT [4] et SURF [2]. Il repose sur le détecteur de points-clés FAST pour la localisation et sur le descripteur binaire BRIEF pour la description, augmenté d'une composante d'orientation afin de garantir l'invariance à la rotation. La comparaison de descripteurs s'effectue via la distance de Hamming, ce qui le rend particulièrement efficace en termes de temps de calcul. Dans le cadre de la détection de quasi-doublons, ORB permet d'établir des correspondances locales entre paires d'images, même en présence de transformations géométriques modérées telles que la rotation ou le recadrage partiel. Toutefois, sa robustesse reste limitée face aux changements d'échelle importants et aux variations d'illumination prononcées. Son principal avantage réside dans son faible coût computationnel, le rendant adapté à un traitement à grande échelle sur des jeux de données de taille importante.

2.1.4 DELF

DELF (DEep Local Features) [5] est une méthode d'extraction de descripteurs locaux basée sur un CNN entraîné sur de grandes collections d'images. Contrairement aux descripteurs classiques tels qu'ORB ou SIFT, DELF apprend des représentations locales directement à partir des données, ce qui lui confère une bien meilleure robustesse aux changements de point de vue, d'échelle et de conditions d'illumination. Un mécanisme d'attention permet de sélectionner automatiquement les points-clés les plus discriminants pour la tâche de mise en correspondance. DELF a initialement été développé pour la reconnaissance visuelle de lieux et s'est imposé comme une méthode de référence dans ce domaine. Son application à la détection de quasi-doublons est naturelle : les correspondances denses et précises qu'il établit entre images permettent de détecter des similarités même sous des transformations complexes. Son principal inconvénient est son coût computationnel élevé, tant pour l'extraction des descripteurs que pour la mise en correspondance.

2.2 Jeu de données

Nous disposons d'un jeu de données fourni par la Fondation Science4Reefs. Ce dernier est composé de 5 967 photographies sous-marines de coraux de Moorea, en Polynésie française. Parmi celles-ci, 633 images réparties en 159 groupes de quasi-doublons ont été identifiés manuellement (cf. Figure 1). Cette quantité d'images reste limitée pour l'entraînement de modèles d'apprentissage profond. En pratique, les architectures utilisées en vision par ordinateur atteignent une meilleure stabilité et une capacité de généralisation satisfaisante à partir de plusieurs dizaines de milliers d'images, voire de centaines de milliers lorsque les variations visuelles sont importantes, comme c'est le cas en milieu sous-marin. Ainsi, ce jeu de données est utilisable dans cette étude comparative, mais ne permet pas d'exploiter pleinement les modèles étudiés sans risque de surapprentissage.



FIGURE 1 – Un exemple de quasi-doublons dans le jeu de données fourni par la fondation Science4Reefs

2.3 Métriques d'évaluation

Les méthodes de détection de quasi-doublons produisent généralement une liste d'images du jeu de données, classées par similarité décroissante avec l'image requête. Pour évaluer ces classements, deux métriques complémentaires sont couramment utilisées dans la littérature : le Recall@K et la mean Average Precision (mAP). Le Recall@K évalue la capacité du système à retrouver les bons résultats, tandis que la mAP mesure la qualité globale du classement. Ainsi, elles permettent une évaluation robuste des méthodes proposées.

Le **Recall@K** mesure la proportion de requêtes pour lesquelles au moins un near-duplicate apparaît parmi les K images les plus similaires retournées par le modèle. Il se calcule comme suit :

$$\text{Recall@K} = \frac{N_{\text{hit}}}{N_q} \tag{1}$$

où N_{hit} est le nombre de requêtes ayant au moins un near-duplicate dans les K images les plus similaires selectionnées par le modèle pour une requête donnée, et N_q le nombre total de requêtes.

La **mAP** évalue à la fois la pertinence et l'ordonnancement des résultats retournés par le système. Une mAP élevée indique que les images les plus similaires à la requête sont bien positionnées en tête du classement. Pour chaque requête, on calcule d'abord l'Average Precision (AP), qui mesure la qualité du classement des éléments pertinents dans la liste :

$$\text{AP} = \frac{1}{N} \sum_{i=1}^{k} P(i) \cdot \text{rel}(i) \tag{2}$$

où N est le nombre total d'éléments pertinents, $P(i)$ est la précision au rang i, et $\text{rel}(i)$ vaut 1 si l'élément au rang i est pertinent, 0 sinon. Cette valeur est ensuite moyennée sur l'ensemble des requêtes :

$$\text{mAP} = \frac{1}{Q} \sum_{q=1}^{Q} \text{AP}_q \tag{3}$$

où Q est le nombre total de requêtes.

De plus, deux métriques ont été considérées pour mieux caractériser le comportement du système lors de l'implémentation du clustering : le Recall et la Pureté.

Le **Recall** reflète la capacité du système à récupérer l'ensemble des quasi-doublons. Une valeur élevée indique que le système minimise les faux négatifs. Il se calcule ainsi :

$$\text{Recall} = \frac{N_{\text{correct}}}{N_{\text{ND}}} \tag{4}$$

où N_{correct} est le nombre de quasi-doublons correctement identifiés, et N_{ND} le nombre total de quasi-doublons dans le jeu de données.

La **Pureté** évalue la qualité des résultats de clustering en mesurant l'homogénéité des groupes générés. Un cluster est considéré pur si toutes ses images appartiennent au même groupe de near-duplicate tel que défini par les annotations manuelles. Cette métrique est utilisée à titre indicatif uniquement, les annotations manuelles n'étant pas considérées comme fiables. Elle se calcule comme suit :

$$\text{Pureté} = \frac{N_{\text{pur}}}{N_{\text{cluster}}} \tag{5}$$

où N_{pur} est le nombre d'images appartenant à des clusters purs, et N_{cluster} le nombre total d'images appartenant à des clusters de taille supérieure à 1.

2.4 Méthodologie

La détection de quasi-doublons est décomposée en cinq étapes séquentielles : l'extraction de descripteurs des images, le calcul d'une matrice de similarité entre paires d'images, la sélection d'un seuil de similarité, et une phase de clustering (cf. Figure 2). Les descripteurs permettent de comparer les images deux à deux via la matrice de similarité, le seuil transforme ces scores continus en décisions binaires similaire/non similaire, et le clustering agrège ces relations locales en groupes cohérents. Chaque étape conditionne directement la suivante, ce qui justifie une évaluation progressive et indépendante de chacune d'elles.

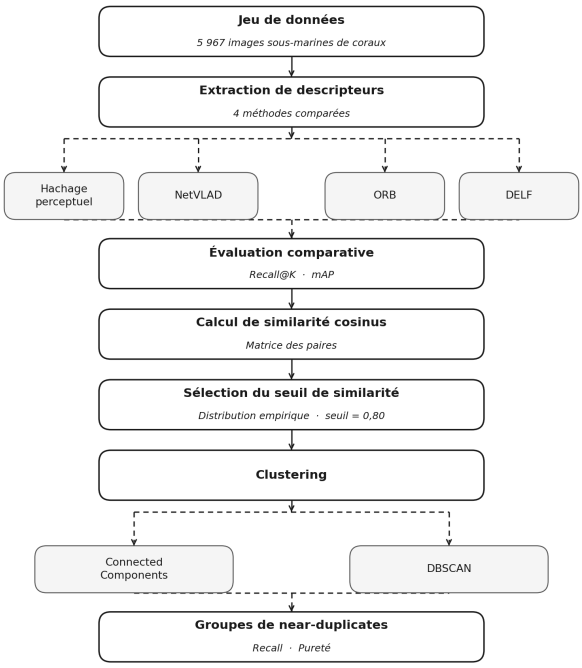


FIGURE 2 – Organigramme résumant étapes principales de la démarche.

3 Étude comparative

Cette section présente les résultats obtenus pour chacune des méthodes d'extraction de caractéristiques, en évaluant leur performance sur le jeu de données d'images de coraux. À partir des descripteurs produits par la méthode ayant obtenu les meilleures résultats, NetVLAD, une chaîne de traitement complète de détection de quasi-doublons est ensuite mise en place. Elle repose sur le calcul d'une matrice de similarité et le regroupement des images via un algorithme de clustering. Deux approches sont comparées (Connected Components et DBSCAN) en évaluant leur performance pour différents seuils de similarité.

3.1 Extractions de caractéristiques

3.1.1 Hachage perceptuel

Plusieurs méthodes de hachage perceptuel issues de la bibliothèque Python ImageHash [3] ont été étudiées. Cette bibliothèque propose différentes implémentations de hachage d'images, conçues pour mesurer la similarité visuelle tout en restant peu coûteuses en calcul. Une série d'expériences comparatives a été menée afin d'évaluer les performances de ces techniques. Les méthodes testées incluent aHash, dHash, pHash, wHash, colorHash et un crop-resistant hash. Cette phase expérimentale a permis de constater que le pHash, bien que largement utilisé dans la littérature, présentait une performance inférieure sur ce cas d'étude par rapport à d'autres méthodes telles que wHash ou aHash. Les performances ont été mesurées à l'aide de la métrique Recall@K, et les résultats obtenus sont résumés dans le tableau 1 :

	R@1	R@5	R@10
aHash	0.42	0.53	0.60
wHash	0.40	0.53	0.60
dHash	0.25	0.36	0.40
colorHash	0.24	0.46	0.56
crop resistant hash	0.21	0.29	0.34
pHash	0.18	0.28	0.34

TABLE 1 – Comparaison des méthodes de hachage perceptuel.

L'analyse qualitative des différentes fonctions met en évidence l'influence du type de représentation visuelle utilisé par les méthodes de hachage. Les approches basées sur l'apparence globale de l'image, comme le aHash, obtiennent de bons résultats car les quasi-doublons du jeu de données présentent des structures très similaires et peu de transformations complexes.
À l'inverse, le pHash, qui repose sur une représentation fréquentielle plus abstraite, perd une partie de l'information discriminante lorsque les différences entre images reposent sur des variations locales ou de légers changements de point de vue. Les méthodes sensibles aux détails locaux, comme le wHash et le dHash, obtiennent des performances intermédiaires.
Enfin, les méthodes basées uniquement sur la couleur ou conçues pour être très robustes aux transformations fortes montrent des résultats plus limités. Le colorHash peut rapprocher des images aux palettes similaires mais de contenu différent, tandis que les approches tolérant fortement les recadrages perdent une partie de leur pouvoir discriminant.
Dans l'ensemble, ces résultats suggèrent que, pour ce jeu de données, des représentations simples capturant directement les similarités visuelles globales ou locales sont plus adaptées que des méthodes plus abstraites ou fortement robustes aux transformations.

Hachage hybride couleur–structure. Afin de dépasser les limites des approches basées sur un seul type de hachage perceptuel, une stratégie de hachage hybride a été explorée. Elle consiste à combiner deux méthodes de hachage complémentaires à l'aide d'une distance pondérée :

$$d = \alpha \, | \, Hash_1^{\text{query}} - Hash_1^{\text{image}} \, | \\ + \beta \, | \, Hash_2^{\text{query}} - Hash_2^{\text{image}} \, | \qquad (6)$$

Les meilleures performances ont été obtenues avec la combinaison colorHash + aHash. Les résultats montrent qu'une pondération majoritaire du colorHash (80 %) améliore significativement les performances, avec un gain supérieur à 10 % sur la métrique Recall@10 par rapport aux méthodes individuelles.
Cette amélioration s'explique par la complémentarité des informations capturées. Le colorHash agit comme un descripteur global permettant de regrouper rapidement les images partageant une palette chromatique similaire, ce qui est particulièrement pertinent pour les images de coraux qui possèdent souvent des signatures de couleur caractéristiques. Toutefois, cette information seule reste insuffisante

aHash (α)	colorHash (β)	R@1	R@5	R@10
0,2	0,8	0,48	0,66	0,73
0,4	0,6	0,47	0,64	0,70
0,6	0,4	0,44	0,59	0,66
0,8	0,2	0,39	0,55	0,62

TABLE 2 – Performance du hachage hybride combinant aHash et colorHash selon différentes pondérations.

car des structures différentes peuvent présenter des couleurs proches. L'intégration du aHash introduit alors une contrainte géométrique simple basée sur la structure globale de l'image, qui agit comme un filtre permettant de réduire les faux positifs générés par la similarité de couleur seule. Ainsi, le colorHash assure une sélection large basée sur l'apparence chromatique, tandis que le aHash joue le rôle de filtre structurel permettant de vérifier la cohérence géométrique des images candidates. Cette complémentarité explique les gains observés et souligne l'intérêt d'associer des méthodes de hachage perceptuel capturant des propriétés visuelles différentes.

3.1.2 NetVLAD

La méthode NetVLAD permet d'agréger des descripteurs locaux d'images en un vecteur global compact. Après l'implémentation de la méthode, un fine-tuning a permis de l'adapter aux similarités fines entre images de coraux proches, tout en conservant les représentations générales apprises sur de larges jeux de données de référence. Le jeu de données annoté étant limité et fortement déséquilibré, l'entraînement complet du modèle n'était pas envisageable. Pour le fine-tuning, une validation croisée à cinq folds, organisée par groupes de quasi-doublons, a été mise en place pour garantir une évaluation plus fiable et reflètant la capacité du modèle à généraliser sur de nouveaux coraux. Les différents hyperparamètres du modèle ont ensuite été ajustés afin de déterminer une configuration optimale. Le backbone choisi pour l'extraction des descripteurs est le CNN ResNet-50, préentraîné sur le jeu de données ImageNet, dont les couches intermédiaires ont été dégelées, permettant un apprentissage approfondi. Deux fonctions de perte ont été comparées : la MultiSimilarityLoss et la HardTripletLoss. Le choix s'est porté sur la MultySimilarityLoss qui fournit des gradients progressifs et favorise une convergence stable en début d'entraînement. La pondération des paires difficiles a permis de rapprocher les quasi-doublons et d'éloigner les images non similaires. Enfin, le taux d'apprentissage a été optimisé grâce à un ordonnanceur et à la mise en place de taux d'apprentissage distincts pour le CNN et pour la couche NetVLAD.

Une fois le fine-tuning réalisé sur les cinq folds de dix epochs, le modèle final a été sélectionné par une évaluation de type Leave-One-Out qui consiste à comparer chaque image requête, représentative d'un groupe de quasi-doublons, à toutes les images du même fold, à l'exception de celles de son propre groupe. Cette approche permet d'évaluer la qualité des descripteurs en évitant les biais liés à une évaluation sur l'ensemble complet des 5 967 images du jeu de données. Cependant, une telle évaluation a permis d'estimer les performances réelles du modèle en situation pratique. Pour cette évaluation, le jeu de données a été séparé en un jeu d'images requêtes et un jeu d'images candidates : le premier contient une image représentative par individu ayant des quasi-doublons, tandis que le second contient toutes les autres images. Cette configuration simule un cas réel de recherche de quasi-doublons et permet de mesurer les performances finales du modèle en termes de mAP et de Recall@k, assurant la généralisation du modèle NetVLAD. Le modèle final retenu correspond à la version du fold 4, epoch 10, ayant fourni, sur l'ensemble du jeu de données, la meilleure mAP (82,66%) et le troisième meilleur Recall@1 (86,79%), ainsi que des Recall@10, et plus, dépassant les 98% de récupération (cf. Figure 3).

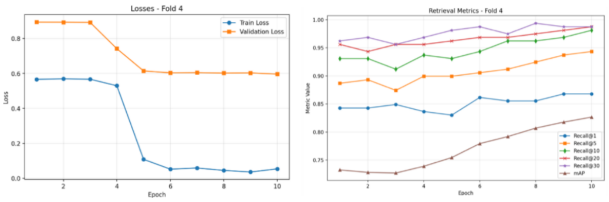


FIGURE 3 – Évolution de la perte en entraînement et en validation (gauche) et des métriques Recall@k et mAP (droite) pour le fold 4 en fonction des epochs

3.1.3 ORB

L'algorithme ORB est particulièrement prisé pour son excellente rapidité d'exécution et son absence de brevets, le rendant idéal pour les applications temps réel. ORB détecte principalement des coins et motifs structurés, et des zones à fort contraste. Sur les images de coraux, les textures sont souvent irrégulières et organiques, avec peu de coins nets, ce qui limite le nombre de point clé pertinents. Cela explique les performances limitées du modèle résumées dans le tableau 3.

R@1	R@5	R@10	R@20	R@30	mAP
0,19	0,38	0,43	0,49	0,55	0,12

TABLE 3 – Performance de la méthode ORB.

3.1.4 DELF

Afin de spécifier l'extraction des données au contexte spécifique des coraux, la piste d'un fine-tuning du modèle DELF a été identifiée comme une évolution pertinente. Cette approche consisterait à spécialiser le CNN pré-entraîné sur des images de récifs coralliens afin d'orienter les mécanismes d'attention vers les structures biologiques d'intérêt. Le fine-tuning de DELF s'est donc appuyé sur le dépôt DeLF-pytorch [7], qui a été cloné et adapté au contexte du projet. Ce dépôt fournit une implémentation modulaire de DELF ce qui permet le fine-tuning du modèle et l'extraction de descripteurs locaux. Deux configurations de fine-tuning ont donc été évaluées : le fine-tuning du backbone uniquement, ainsi que le fine-tuning conjoint du backbone et du

module de points-clés. Les hyperparamètres ont été ajustés entre les différents entraînements réalisés, et les résultats principaux sont résumés dans le tableau 4.

Méthode	R@1	R@5	R@10	mAP
Zéro-shot	0,73	0,79	0,82	0,25
Backbone, LR=1e-4	0,53	0,64	0,68	0,36
Backbone, LR=5e-5	0,30	0,44	0,48	0,27
Backbone, LR=1e-5	0,43	0,53	0,57	0,36
Full, LR=1e-4	0,79	0,87	0,90	0,59
Full, LR=5e-5	0,80	0,87	0,90	0,60
Full, LR=1e-5	0,81	0,88	0,90	0,61

TABLE 4 – Résultats du fine-tuning de DELF selon les méthodes et les paramètres.

Dans le cas du fine-tuning du backbone uniquement, les performances chutent par rapport au modèle zero-shot, quel que soit le taux d'apprentissage. Cela montre que l'adaptation des poids du CNN seule perturbe les représentations visuelles apprises lors du premier entraînement, sans permettre une réorganisation cohérente des points-clés. Le module d'attention, qui reste figé, continue de sélectionner des régions selon des critères inadaptés aux nouvelles caractéristiques extraites par le backbone. Cela crée un désalignement qui conduit à des descripteurs moins discriminants, ce qui explique la baisse des métriques. À l'inverse, le fine-tuning conjoint du backbone et du module d'attention conduit à des gains significatifs par rapport à la version zéro-shot. Cela s'explique par la capacité du modèle à adapter en même temps l'extraction des caractéristiques locales et le mécanisme d'attention qui sélectionne les régions pertinentes. Le module d'attention apprend ainsi à se concentrer sur des zones plus distinctives du corail comme les textures, plutôt que sur des éléments parasites ou trop génériques.

3.2 Regroupement des images

La méthode NetVLAD ayant fourni les meilleurs résultats, les descripteurs qu'elle a généré sont utilisés lors des étapes suivantes.

Calcul de similarité. Après l'extraction des descripteurs de chaque image du jeu de données, on calcule la similarité entre chaque paire de descripteurs pour déterminer si deux images sont des quasi-doublons. Cette comparaison repose sur la similarité cosinus afin de leur attribuer un score de similarité. Elle se calcule ainsi :

$$\text{Similarité cosinus}(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} \qquad (7)$$

Deux images sont des quasi-doublons quand leur score de similarité dépasse un certain seuil. Une méthodologie s'appuyant sur les scores de similarité va alors permettre de le déterminer. Tout d'abord, les similarités entre toutes les paires d'images sont calculées afin de constituer une matrice de similarité. La mise en parallèle du jeu de données manuellement annoté avec les distributions de similarité (cf. Figure 4) révèle un score moyen de 0,7 pour les quasi-doublons et de 0,27 pour les singletons, ce qui confirme que le modèle distingue nettement les deux groupes. Cependant, un chevauchement subsiste entre 0,4 et 0,7. Plusieurs seuils ont donc été testés dans le cadre du clustering afin de déterminer un seuil satisfaisant.

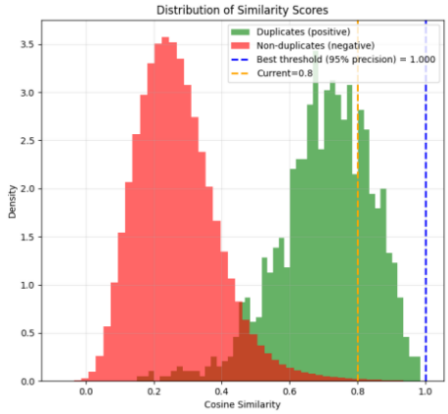


FIGURE 4 – Distribution des scores de similarité par paire d'images selon leur statut dans le jeu de données annoté : near-duplicate (vert) ou singleton (rouge)

Clustering. La phase de clustering est essentielle pour passer de la comparaison de similarité entre deux images à la détection de groupes de quasi-doublons, c'est-à-dire des clusters. Deux méthodes principales ont été étudiées : la méthode de chaînage transitif Connected Components (composantes connexes) et l'algorithme de partitionnement Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Connected Components forme un graphe de similarité non orienté, où chaque nœud représente une image et une arête relie deux images si leur score de similarité dépasse le seuil fixé. La méthode DBSCAN, quant à elle, regroupe automatiquement les images à partir de leur score de similarité en prenant en compte le seuil fixé, et un nombre d'image minimal pour former un groupe de quasi-doublons. Par expérimentation, la méthode retenue qui fournit de meilleurs résultats est Connected Components (cf. Figure 5).

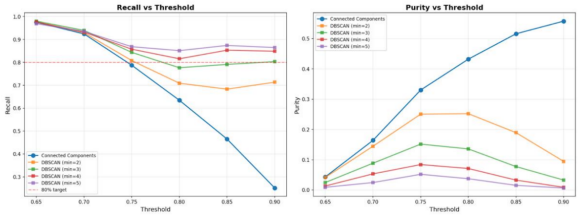


FIGURE 5 – Comparaison des approches Connected Components (bleu) et DBSCAN selon le Recall et la Pureté en fonction du seuil de similarité

Le choix du seuil a été guidé par un compromis : un seuil trop faible (0,75) génère des méga-clusters, tandis qu'un seuil trop strict (0,85) exclut de nombreux quasi-doublons.

Le seuil de 0,80 a été retenu, offrant un équilibre entre Recall (63%) et Pureté (43%). Cette valeur a été choisie de manière conservatrice afin de privilégier la précision et limiter les faux positifs.

4 Discussion

Cette étude s'est articulée autour d'une analyse exploratoire de publications de détection de quasi-doublons et de reconnaissance visuelle de lieux. Elle a permis d'identifier quatre approches pertinentes adaptées au milieu sous-marin et aux motifs complexes des coraux.

Les expérimentations, menées sur le jeu de données fourni par la Fondation Science4Reefs, montrent que la méthode ORB est peu adaptée aux besoins de cette étude, tandis que le hachage perceptuel et DELF offrent des performances satisfaisantes, indiquant un potentiel d'adaptation intéressant. Le gain significatif obtenu par le fine-tuning complet de DELF souligne que l'adaptation au contexte spécifique des images est essentielle pour la détection de quasi-doublons, qui s'appuie sur des motifs fins caractéristiques. NetVLAD, quant à elle, a surpassé les autres approches, atteignant un Recall@1 de 86,79% et une mAP de 82,66%. Ces résultats démontrent que, correctement fine-tunée, la méthode NetVLAD constitue un outil efficace pour détecter des quasi-doublons dans des images de coraux. La phase de clustering repose sur la méthode Connected Components et un compromis conservateur entre récupération et précision, pouvant conduire à la formation de méga-clusters. Étant donné que le clustering minimise les faux positifs, l'intégration d'une validation humaine peut s'avérer pertinente pour garantir la qualité des clusters.

La chaîne de traitement développée permet un nettoyage fiable du jeu de données, réduisant les besoins en stockage et en calcul, tout en ouvrant d'autres perspectives d'utilisation, telles que la détection d'images d'un même individu de corail, voire d'espèces spécifiques. Les pistes de poursuite de cette étude concernent notamment l'expérimentation de nouvelles méthodes d'extraction de caractéristiques d'images et de techniques de clustering afin d'améliorer les performances. Par ailleurs, le jeu de données fourni reste limité en nombre d'images, ce qui restreint le fine-tuning des méthodes de deep learning et leur capacité à apprendre des caractéristiques spécifiques aux coraux. Cette contrainte a guidé des choix méthodologiques pragmatiques, mais souligne l'importance de poursuivre ces travaux sur des jeux de données plus vastes et diversifiés afin d'augmenter la généralisation des modèles et leur robustesse en contexte réel. En effet, la méthodologie proposée, bien qu'appliquée ici à des images de coraux de Moorea, présente un potentiel de généralisation à des images de coraux issues d'autres régions géographiques, et à d'autres domaines caractérisés par des acquisitions séquentielles (faune marine, biodiversité terrestre). Cette généralisation est conditionnée à une adaptation au domaine cible, notamment concernant le paramétrage des modèles, selon la finesse des motifs à détecter.

5 Conclusion

L'objectif de cette étude est de comparer différentes méthodes de détection de quasi-doublons, appliquées à un jeu de données d'images de coraux, caractérisé par des motifs fins et les variations naturelles de l'environnement aquatique. Cette étude a permis d'adapter, d'évaluer et de comparer quatre approches d'extraction de caractéristiques d'images : le hachage perceptuel, les descripteurs globaux (NetVLAD), et les descripteurs locaux reposant sur la détection de points-clés (ORB et DELF). Elle a également permis de développer une méthode de clustering, complétant alors la chaîne de traitement de détection de quasi-doublons. Les travaux réalisés apportent ainsi une étude comparative détaillée de différentes méthodes de détection de quasi-doublons, adaptées au domaine corallien sous-marin, encore peu documenté en vision par ordinateur. Ainsi, ces travaux ouvrent la voie à des applications de suivi et de préservation de récifs coralliens.

Remerciements

Nous tenons, tout d'abord, à remercier Marie-Hélène Abel, professeure à l'UTC et membre de l'UMR CNRS Heudiasyc, pour sa grande disponibilité et la qualité de ses conseils. Nous tenons également à remercier Arjen Kraneveld, référent technique, et Lionel Mathieu, référent gestion de projet, au sein de la cellule d'innovation Lab'Innov de l'entreprise Aubay, pour leur accompagnement tout au long du développement de ces travaux. Enfin, nous tenons à remercier Laetitia Hédouin, directice de recherche au CNRS et présidente de la Fondation Science4Reefs, pour la qualité de nos échanges et pour son engagement dans ce projet dédié à la préservation des récifs coralliens.

Références

[1] Relja Arandjelovic , Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad : Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF : Speeded up robust features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006.

[3] Johannes Buchner. ImageHash : A python perceptual image hashing module, 2013.

[4] David G. Lowe. Distinctive image features from scale-invariant keypoints. volume 60, pages 91–110. Springer, 2004.

[5] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3456–3465. IEEE, 2017.

[6] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB : An efficient alternative to SIFT

or SURF. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE, 2011.

[7] Minchul Shin. DeLF-pytorch : Deep local features in PyTorch, 2021. Dépôt GitHub, consulté en 2025.

[8] Christoph Zauner. Implementation and benchmarking of perceptual image hash functions. In *Media Watermarking, Security, and Forensics III*, volume 7880. SPIE, 2010.

The attention illusion: overcoming the localization-prediction discrepancy in biomass ViT-based regression

Farida Nchare[1,2,3], Norbert Tsopze[2,3], Corina Iovan[1]

[1] Institut de Recherche pour le Développement, UMR ENTROPIE, Noumea, BP A5 - 98848 cedex, New Caledonia
[2] Department of Computer Science, University of Yaounde I, Yaounde, P.O. Box 812, Cameroon
[3] Sorbonne University, IRD, UMMISCO, F-93143, Bondy, France

Abstract

Biomass estimation from images is important for ecological monitoring but relies on morphometric features that are not always visually salient. We observe that while Vision Transformers (ViTs) produce attention maps suggesting precise spatial localization, this visual focus does not guarantee accurate weight prediction. Using a frozen DINO-based ViT for bivalve biomass estimation, we show that morphometric information is primarily encoded in intermediate layers rather than final semantic representations. Applying learnable multi-head attention pooling over these layers significantly improves regression performance, achieving an $R^2 = 0.950$ and reducing RMSE by 39.6% over standard CLS aggregation. These results reveal an attention illusion in ViTs and demonstrate the critical role of intermediate features for high-precision morphometric regression.

Keywords

Vision Transformers, Attention analysis, Feature aggregation, Biomass estimation, Explainability.

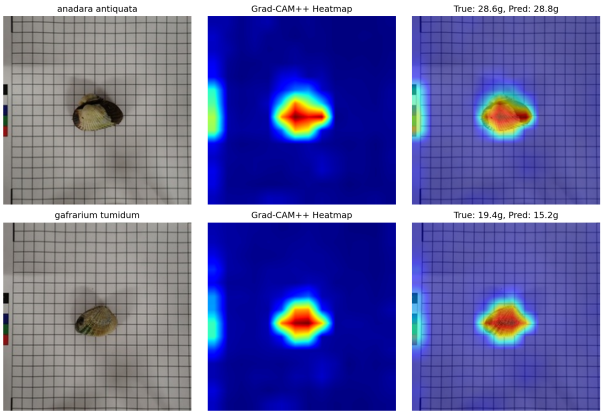


Figure 1: DINO self-attention maps for two bivalve specimens from different species. The model predicts 28.8 g for a 28.6 g specimen (top) but 15.2 g for a 19.4 g specimen (bottom). Despite similar localization patterns, prediction accuracy differs significantly, illustrating the *attention illusion*.

1 Introduction

Estimating biomass directly from images enables non-invasive monitoring of marine species populations across ecosystems. Unlike classification tasks that focus on categorical identity, biomass estimation relies on *morphometric information*, meaning structural properties such as size, curvature, and volumetric features that correlate with organism mass. Extracting these continuous geometric signals from images is challenging, particularly when specimens exhibit strong cross-species variability. Individuals with similar visual appearance may differ in density, making accurate mass estimation dependent on subtle morphological features.

Vision Transformers (ViTs) trained with self-supervised objectives such as DINO [3] produce attention maps that sharply delineate object boundaries without task-specific supervision. This spatial precision suggests that such models should capture the geometric structure required for morphometric estimation. However, standard pipelines extract the global representation from the final-layer `CLS` token, a design inherited from classification models [5]. Classification and regression impose fundamentally different representational requirements. DINO encourages semantic invariance by training the `CLS` token to produce similar embeddings for augmented views of the same image, preserving object identity while discarding variations in scale or geometry. In contrast, biomass regression requires these variations to remain encoded in the representation. As transformer layers progressively transform low-level features into higher-level semantics [9], fine-grained morphometric information may therefore be compressed or lost. Consequently, as illustrated in Figure 1, two specimens can exhibit nearly identical attention patterns while producing very different prediction accuracy. We refer to this mismatch between spatial localization and predictive performance as the *attention illusion*.

This work investigates where morphometric information is encoded within the ViT hierarchy. Through layer-wise regression analysis, we show that predictive performance peaks at intermediate transformer layers rather than at the final semantic layer. Motivated by this observation,

we propose Adaptive Multi-head Attention Pooling (A-MAP), a feature aggregation strategy that operates on patch tokens from these intermediate layers instead of relying on the global CLS representation. A-MAP builds upon MAP [7] and uses learnable attention to aggregate spatial features, enabling different heads to capture complementary geometric structures. To further enrich the representation, the method aggregates tokens from adjacent intermediate layers, providing a multi-level description of specimen geometry.

We evaluate the proposed approach on a dataset of more than 2000 images of bivalves from two species (*Anadara antiquata* and *Gafrarium tumidum*) with weights ranging from 0.2 g to 66.6 g. Using a frozen DINO ViT-S/16 backbone, the model is trained without species labels so that predictions must rely on morphometric cues rather than species identity. Our analysis reveals that morphometric information peaks at intermediate layers, and aggregating these representations with A-MAP achieves $R^2 = 0.950$, reducing RMSE by 39.6% compared to standard CLS aggregation.

Our contributions are as follows:

- We identify the *attention illusion* in ViT-based regression, showing that accurate spatial attention does not guarantee accurate morphometric prediction.
- We conduct a systematic layer-wise analysis demonstrating that morphometric information peaks at intermediate transformer layers rather than at the final semantic representation.
- We propose A-MAP, an attention-based pooling strategy that aggregates intermediate-layer patch tokens and improves by $+10.0\%$ regression performance over standard CLS aggregation.

2 Related Work

ViT representation hierarchy. Vision Transformers (ViTs) exhibit a progressive evolution of representations across their depth, transitioning from localized spatial patterns to abstract semantic descriptors [9, 5]. Early layers retain fine-grained textures and edge information, whereas deeper layers increasingly encode class-level abstractions [3]. Self-supervised frameworks such as DINO demonstrate that attention maps can highlight coherent object regions without explicit supervision [3]. Despite this, most downstream applications still rely exclusively on the final-layer [CLS] token as a global image representation [5]. Recent studies suggest that spatial relationships and object structure are distributed across the transformer feature space rather than concentrated in a single token [8]. However, the representational hierarchy of ViTs is largely shaped by classification objectives. As a result, the class-discriminative information emphasized in the deepest layers may not preserve the continuous spatial structure required for precise morphometric regression.

Feature aggregation beyond CLS. To address the limitations of the [CLS] token, several aggregation strategies have been proposed to better exploit patch tokens. Multi-head Attention Pooling (MAP) introduces learnable queries that attend to spatial tokens, enabling a more flexible representation than fixed global pooling [7]. Other approaches focus on multi-scale feature fusion. For instance, HAFA [4] and SDPT [2] combine representations from multiple transformer layers to capture both low-level textures and high-level semantic features. In regression tasks, architectures such as FASNet [11] have successfully employed shallow and deep features jointly for crowd density estimation. Nevertheless, these multi-scale strategies typically assume that aggregating more features leads to better representations, often performing uniform fusion across layers. Such aggregation may dilute task-relevant information, as signals useful for morphometric estimation could be overlooked by the increasing semantic abstraction of deeper layers. This suggests the need for more selective strategies that explicitly target layers where morphometric features are most informative.

Morphometric regression from images. Image-based estimation of biomass or weight has become an important tool for ecological monitoring [13]. Prior work has developed effective pipelines for fish biomass estimation [12], segmentation-based morphometric analysis [6], and joint detection–estimation frameworks for species-specific prediction [10]. However, these methods typically extract features from the final backbone layer. Whether applied to fish or marine invertebrates [1], they implicitly assume that the deepest representation is optimal for regression. Yet this assumption has rarely been examined for morphometric estimation. This work addresses this gap by investigating the intermediate feature space of ViTs and proposing a regression pipeline designed to extract morphometric information from these layers.

3 Method

We investigate where morphometric information resides in frozen Vision Transformer (ViT) representations and how to exploit it for regression. The method consists of three stages: (1) problem formulation, (2) layer selection via layer-wise regression evaluation, and (3) feature aggregation using Adaptive Multi-head Attention Pooling (A-MAP).

3.1 Problem Setup

Given an RGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the goal is to predict a scalar weight $\hat{y} \in \mathbb{R}^+$ (in grams). We use a frozen DINO ViT-S/16 backbone [3]. For each image, the transformer produces patch token embeddings at each layer l:

$$\mathbf{X}_l \in \mathbb{R}^{N\times D},$$

where $N = 196$ is the number of patch tokens, $D = 384$ the embedding dimension, and $l \in \{0, \ldots, L-1\}$ with $L = 12$ layers. The backbone remains frozen throughout training. Only the pooling and regression layers are optimized.

Standard practice extracts the CLS token from the final layer as a global representation. However, the *attention*

illusion suggests that late-layer representations may discard the morphometric information needed to predict mass. Our objective is therefore to learn a pooling function

$$f : \mathbf{X}_l \to \mathbf{z}$$

that extracts a fixed-size representation $\mathbf{z} \in \mathbb{R}^D$, which is then mapped to the scalar prediction $\hat{y}$ by a regression head. The central question is twofold: at which depth l does the backbone encode the geometric features most useful for mass regression, and how should we aggregate patch tokens to exploit them?

3.2 Layer Selection Protocol

For tasks that depend on geometric features, intermediate representations may preserve more task-relevant information than the final semantic layer. To test whether this representational mismatch reflects a depth-dependent phenomenon, we evaluate regression performance independently at each transformer layer. For each layer $l \in \{0, \ldots, L-1\}$, we extract the patch token representation $\mathbf{X}_l$ and train an independent regression head g_l. Each model uses the same pooling and regression architecture to ensure fair comparison across layers. Validation performance is then used to identify the optimal layer

$$l^* = \arg\max_l R^2(g_l(\mathbf{X}_l)).$$

This procedure is architecture-agnostic with respect to the regression head, as any pooling mechanism can serve as g_l. The specific head used in our analysis is described in Section 4.

3.3 Multi-head Attention Pooling

Multi-head Attention Pooling (MAP) [7] replaces the fixed `CLS` token with a learnable aggregation over patch tokens. A learnable query attends to the token sequence using standard multi-head scaled dot-product attention, producing a pooled representation in $\mathbb{R}^D$. This mechanism allows the model to emphasize informative spatial regions rather than relying on the fixed `CLS` representation.

While MAP provides effective features aggregation, it operates on tokens from a single layer. The layer selection protocol may reveal a zone of high performance around the peak, suggesting that combining adjacent layers could capture geometric features at complementary levels of abstraction. A-MAP extends MAP to exploit this structure.

3.4 Adaptive Multi-head Attention Pooling (A-MAP)

To exploit the representational structure revealed by the layer selection protocol, A-MAP attends jointly over patch tokens from multiple intermediate layers centered on the morphometric peak. We denote H the number of attention heads and $d_k = D/H$ the per-head dimension. A-MAP operates through two mechanisms: multi-layer concatenation and a peak-based attention bias.

Multi-layer concatenation. Let l_1, l_2, l_3 denote three layers selected by the protocol. This window captures the morphometric peak and its immediate neighborhood, balancing representation richness against computational cost. We concatenate rather than fuse their representations, as concatenation preserves each layer's token identity and lets the attention mechanism learn which depth is most informative per spatial location:

$$\mathbf{X}_{\text{concat}} = [\mathbf{X}_{l_1}; \mathbf{X}_{l_2}; \mathbf{X}_{l_3}] \in \mathbb{R}^{3N \times D} \tag{1}$$

where $[;]$ denotes row-wise concatenation. This yields $3N$ tokens of dimension D, forming a single sequence that spans three levels of the representational hierarchy.

The remaining question is how to bias attention toward the most informative layer without hard-coding the selection.

Attention with peak-based bias. A learnable query $\mathbf{q}_h \in \mathbb{R}^{1 \times d_k}$ for each head h attends over keys $\mathbf{k}_j \in \mathbb{R}^{1 \times d_k}$ derived from $\mathbf{X}_{\text{concat}}$. We augment the standard scaled dot-product score with a learnable per-layer bias:

$$e_j = \frac{\mathbf{q}_h \, \mathbf{k}_j^\top}{\sqrt{d_k}} + b_{l(j)} \tag{2}$$

where $l(j)$ denotes the source layer of token j in the concatenated sequence. The term $b_{l(j)} \in \mathbb{R}$ is a learnable scalar bias assigned to all tokens from layer $l(j)$. The bias for the peak layer l^* is initialized to $+2.0$, while the others are initialized to 0.0. This initialization encourages early attention toward the morphometrically optimal layer while remaining fully learnable during training.

The per-head attended output is computed as:

$$\mathbf{z}_h = \text{softmax}(\mathbf{e})\, \mathbf{V}_h \in \mathbb{R}^{1 \times d_k} \tag{3}$$

where $\mathbf{V}_h \in \mathbb{R}^{3N \times d_k}$ contains the value projections of $\mathbf{X}_{\text{concat}}$ for head h. Each head produces a weighted summary of the most morphometrically relevant tokens across all three layers. The H head outputs are concatenated and projected to $\mathbb{R}^D$ via a linear layer.

The complete pipeline is summarized as:

$$\hat{y} = g\big(\text{A-MAP}\big([\mathbf{X}_{l_1}; \mathbf{X}_{l_2}; \mathbf{X}_{l_3}]\big)\big)$$

3.5 Regression Head

The pooled representation $\mathbf{z} \in \mathbb{R}^D$ is passed to a lightweight MLP that outputs the scalar prediction $\hat{y}$. The pooling module and regression head are trained jointly, while the backbone remains frozen. Implementation details are provided in Section 4.

4 Experiments

We evaluate whether morphometric information required for biomass estimation is preserved in the standard Vision Transformer pipeline. Our experiments address three questions: (1) where geometric information is encoded within the ViT hierarchy, (2) whether the commonly used `CLS` representation retains these features, and (3) whether spatial aggregation over patch tokens can recover them. Using a frozen DINO ViT-S/16 backbone, we perform a layer-wise regression analysis and compare multiple

feature aggregation strategies. The analysis reveals a clear morphometric peak at intermediate layers where regression accuracy is substantially higher than when using the final-layer CLS token. Motivated by this observation, we evaluate Adaptive Multi-head Attention Pooling (A-MAP), which aggregates intermediate patch representations using learnable attention.

4.1 Experimental Setup

Dataset. We evaluate on the *Bivalves* dataset, comprising 2406 RGB images of marine bivalves from three species: *Anadara antiquata*, *Gafrarium tumidum*, and *Gafrarium pectinatum*. Individual weights range from 0.2 g to 66.6 g. We adopt a cross-species evaluation with model trained and tested on the data regardless of the species identity.

Evaluation Metrics. We report four regression metrics: coefficient of determination (R^2), root mean squared error (RMSE, in grams), mean absolute error (MAE, in grams), and mean absolute percentage error (MAPE). All metrics are reported with 95% bootstrap confidence intervals ($N = 1000$ resamples).

Implementation Details. We use a frozen DINO ViT-S/16 backbone with 12 transformer layers, producing 196 patch tokens of dimension 384 per image. The regression head consists of three fully connected layers ($384 \to 256 \to 128 \to 1$) with dropout. We train with AdamW (lr $= 0.001$, wd $= 0.01$) and a ReduceLROnPlateau scheduler (factor $= 0.8$, patience $= 10$). The loss function is MSE. Batch size is 16, maximum epochs 50, with early stopping (patience $= 15$). All experiments use seed 42.

4.2 Locating Morphometric Information in ViT Representations

Before proposing a new aggregation strategy, we investigate where morphometric information is encoded within the ViT hierarchy. We extract patch tokens from each transformer layer and evaluate regression performance using MAP pooling. Figure 2 shows the resulting R^2 curve across layers.

The curve exhibits a clear inverted U-shape with three regimes. Layer 0 corresponds to raw patch embeddings and yields limited predictive performance ($R^2 = 0.593$). Performance rises rapidly in early layers as contour and structural features emerge.

Regression accuracy peaks in intermediate layers (3–6), with the maximum at layer 5 and nearly identical performance at layer 6. We refer to this region as the *morphometric peak*, where geometric descriptors such as shell curvature and growth patterns are best preserved.

Beyond this point performance gradually declines, reflecting the increasing semantic abstraction induced by the DINO objective [3]. The geometric detail required for mass estimation therefore resides primarily in intermediate representations.

Two conclusions follow. First, the standard CLS pipeline relies on representations where morphometric information has already been partially compressed. Second, intermediate patch tokens retain geometric features that can be exploited by spatial aggregation methods.

4.3 Main Results

Table 1 compares different aggregation strategies using the same frozen DINO ViT-S/16 backbone. Results show that performance depends on both the aggregation mechanism and the representation depth.

The CLS baseline extracts a single global vector from the final layer, discarding spatial structure. Adding attention weighting (CLS-Att.) or concatenating CLS tokens from multiple layers (CLS Multiscale) yields moderate improvements but remains limited.

Replacing fixed aggregation with learnable attention produces a larger gain. MAP pooling introduces a learnable query that attends over patch tokens, improving R^2 from 0.864 to 0.926 on the same layer 11 features.

A further improvement arises from selecting the appropriate representation depth. Layer 5 MAP outperforms layer 11 MAP by +2.1 points despite identical pooling. Selecting the appropriate representation depth therefore has a larger impact on morphometric regression performance than the choice of pooling mechanism.

A-MAP combines both insights by applying multi-head attention over patch tokens concatenated from layers $[4, 5, 6]$, centered on the morphometric peak. This achieves the best performance across all metrics, reducing RMSE by 39.6% compared to the CLS baseline.

4.4 Visual Analysis of Predictions

Figure 3 shows predicted versus ground-truth biomass for the CLS baseline and A-MAP. A-MAP predictions cluster closely around the identity line across the full weight range, while the CLS baseline shows systematic underestimation for heavier specimens. Preserving spatial morphometric information from intermediate layers therefore reduces bias, particularly for high-mass individuals where geometric features are most discriminative.

This observation relates to the *attention illusion* introduced in Section 1. Although DINO self-attention precisely localizes the shell contour (Figure 1), CLS-based regression performs substantially worse. Accurate spatial attention does not guarantee that the aggregated representation preserves morphometric information. The CLS token compresses patch features into a single vector optimized for the pretraining objective, causing geometric features useful for regression to be partially lost.

Figure 4 visualizes the spatial attention patterns of individual heads within the A-MAP pooling module. Each head attends to different shell regions, confirming that multi-head attention captures complementary morphometric features.

4.5 Ablation Studies

Effect of Aggregation Strategy. Table 2 compares aggregation strategies on identical layer 11 features. Mean pooling performs worst, indicating that uniform averaging discards important spatial information. Attention-weighted pooling improves performance but remains limited by

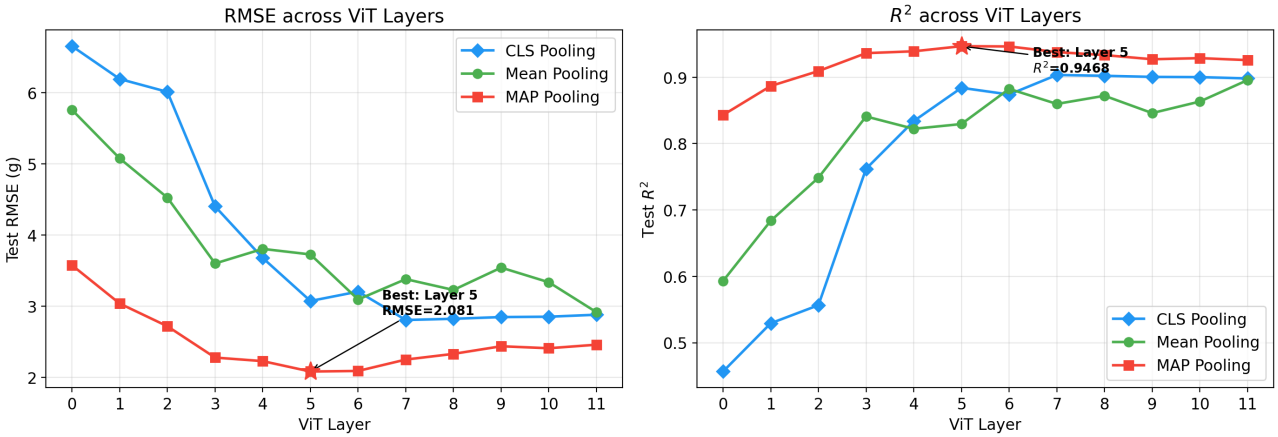


Figure 2: R^2 curve across all DINO ViT-S layers using MAP pooling. The morphometric peak occurs at layer 5 ($R^2 = 0.947$), indicated by ★. Early layers (0–2) encode low-level features; intermediate layers (3–6) encode geometric structure; late layers (7–11) show semantic compression.

Table 1: Comparison of pooling methods on the Bivalves dataset (test split). Best results in **bold**. All metrics reported with 95% bootstrap confidence intervals ($N = 1000$).

Method	Aggregation	Layers	R^2	RMSE (g)	MAE (g)	MAPE (%)
CLS (baseline)	CLS	11	0.864 ± 0.014	3.33 ± 0.10	2.51 ± 0.07	15.0 ± 0.7
CLS-Att.	CLS-Att.	11	0.910 ± 0.009	2.70 ± 0.08	2.03 ± 0.06	11.8 ± 0.6
CLS Multiscale	CLS	[5,6,7,8]	0.910 ± 0.009	2.71 ± 0.12	2.05 ± 0.08	12.0 ± 0.5
MAP	MAP	11	0.926 ± 0.008	2.46 ± 0.08	1.81 ± 0.06	10.4 ± 0.5
Multiscale MAP	MAP	[5,6,7,8]	0.943 ± 0.006	2.15 ± 0.07	1.63 ± 0.06	10.0 ± 0.5
Layer 5 MAP	MAP	5	0.947 ± 0.005	2.08 ± 0.08	1.57 ± 0.06	9.7 ± 0.5
A-MAP (ours)	**A-MAP**	**[4,5,6]**	$\mathbf{0.950 \pm 0.005}$	$\mathbf{2.01 \pm 0.09}$	$\mathbf{1.50 \pm 0.06}$	$\mathbf{8.6 \pm 0.4}$

fixed weights. The CLS-Attention hybrid further gains by incorporating spatial features alongside the `CLS` token. MAP pooling yields the best result, confirming that learnable task-specific attention provides the most effective aggregation.

Table 2: Ablation: aggregation strategy (Layer 11).

Method	R^2	RMSE (g)
Mean	0.826 ± 0.015	3.76 ± 0.17
Att-Weighted	0.879 ± 0.013	3.14 ± 0.17
CLS	0.864 ± 0.014	3.33 ± 0.10
CLS-Att.	0.910 ± 0.009	2.70 ± 0.08
MAP	0.926 ± 0.008	2.46 ± 0.08

Effect of Representation Depth. Table 3 evaluates MAP pooling at different depths. Performance peaks at layer 5, with nearly identical performance at layer 6. Intermediate layers outperform late layers by more than two R^2 points, confirming that representation depth plays a critical role in morphometric regression. This result aligns with the layer-wise analysis in Figure 2, which shows that geometric information is concentrated in intermediate transformer representations [3, 9].

Effect of Multi-layer Fusion. Table 4 compares multi-layer strategies. CLS Multiscale on $[5, 6, 7, 8]$ concatenates `CLS` tokens from multiple layers but achieves only $R^2 = 0.910$, comparable to the single-layer CLS-Att. baseline. Replacing `CLS` concatenation with MAP pooling on the same layers produces a clear gain (+4%), confirming that learnable attention is essential for exploiting multi-layer features. A-MAP narrows the layer window to $[4, 5, 6]$, centered on the morphometric peak, and achieves the best result.

Table 3: Ablation: representation depth (MAP, single layer).

Layer	R^2	RMSE (g)
11	0.926 ± 0.008	2.46 ± 0.08
8	0.933 ± 0.006	2.34 ± 0.08
6	0.946 ± 0.005	2.10 ± 0.08
5	$\mathbf{0.947 \pm 0.005}$	$\mathbf{2.08 \pm 0.08}$

Table 4: Ablation: multi-layer fusion strategy.

Method	Layers	R^2	RMSE (g)
CLS Multiscale	[5,6,7,8]	0.910 ± 0.009	2.71 ± 0.12
Multiscale MAP	[5,6,7,8]	0.943 ± 0.006	2.15 ± 0.07
A-MAP (ours)	**[4,5,6]**	$\mathbf{0.950 \pm 0.005}$	$\mathbf{2.01 \pm 0.09}$

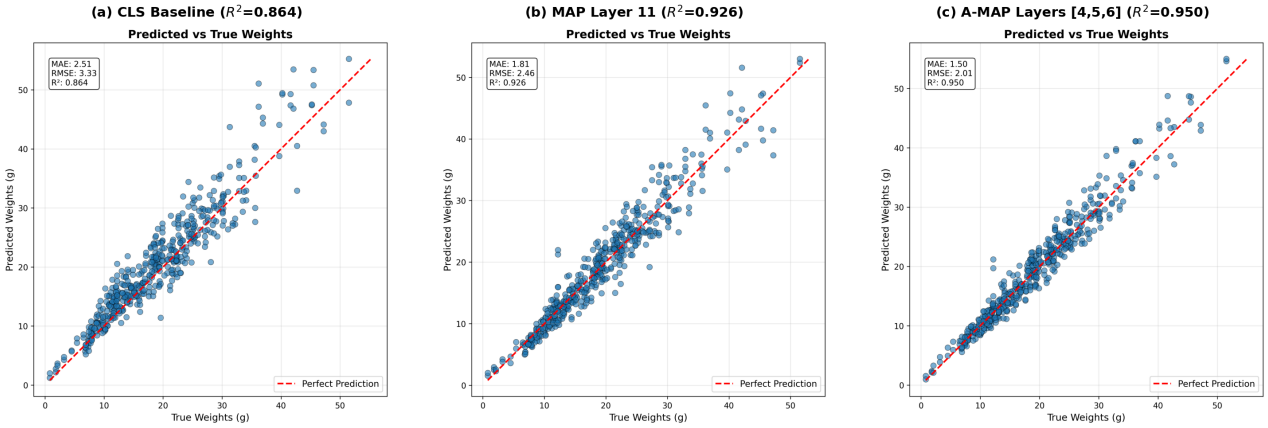


Figure 3: Predicted vs. actual weight scatter plots for the `CLS` baseline (Layer 11, top row) and A-MAP (bottom row). Columns show all species, *Anadara antiquata*, and *Gafrarium* spp., respectively. A-MAP predictions align closely with the identity line across the full weight range.

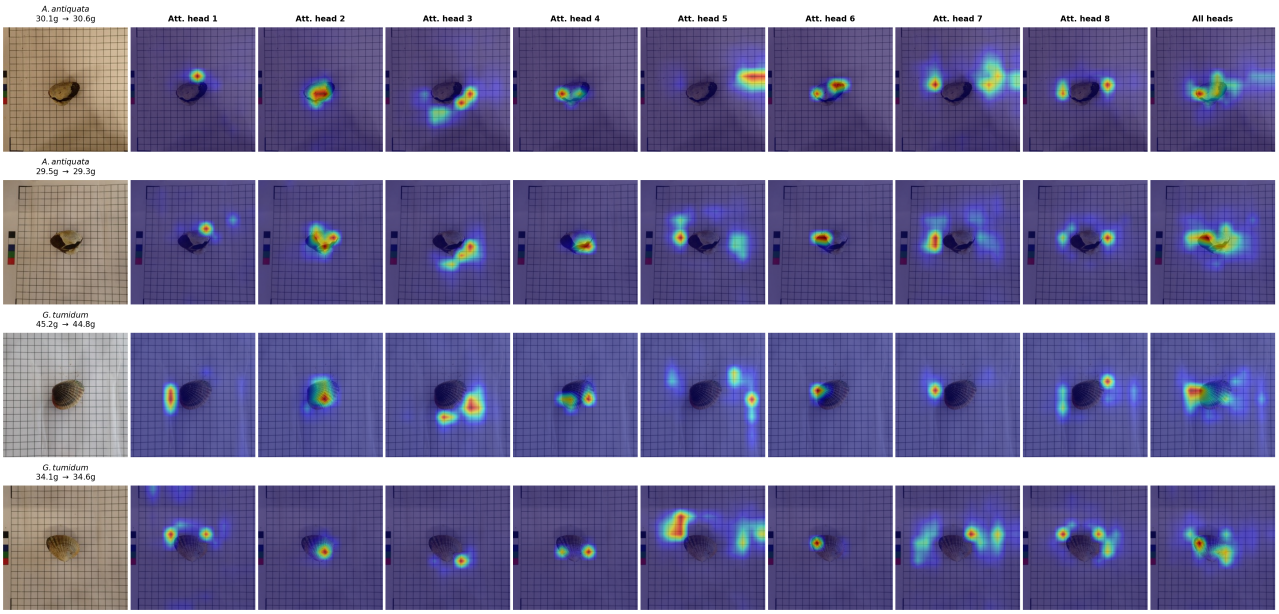


Figure 4: Spatial attention maps from the 8 attention heads of the A-MAP pooling module for two *Anadara antiquata* (top) and two *Gafrarium tumidum* (bottom) specimens. Each column corresponds to one learnable attention head attending over patch tokens from layers $[4, 5, 6]$. The rightmost column shows the aggregated attention across all heads. Different heads specialize in distinct morphometric structures.

4.6 Limitations

- We evaluate on a single morphological group (marine bivalves); generalization to other species or object categories remains untested.
- Our analysis uses a single backbone (DINO ViT-S/16); the morphometric peak layer may shift across architectures or pretraining objectives.
- All experiments use a frozen encoder; fine-tuning the backbone could redistribute morphometric information across layers.
- The dataset size is moderate (2406 images), which may limit the stability of deeper regression architectures.

5 Conclusion

In this work, we identified an *attention illusion* in DINO ViT: self-attention maps precisely localize objects, yet the CLS representation discards the geometric detail needed for weight regression. A systematic layer-wise analysis reveals that morphometric information peaks at intermediate depth (layer 5), not at the final semantic layer. A-MAP, a

learnable attention pooling over patch tokens from layers $[4, 5, 6]$ with peak-centric bias, achieves $R^2 = 0.950$. Compared to the CLS baseline, this reduces RMSE by 39.6% (3.33 g to 2.01 g) and MAPE by 42.7% (15.0% to 8.6%). Combining intermediate geometric features with late semantic context via gated fusion and validating across other physical property regression tasks remain promising directions for future work.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011014241 made by GENCI.

Data availability

The dataset used during the current study is available in the public repository Zenodo 10.5281/zenodo.17936025.

References

[1] Johanna Ärje, Claus Melvad, Mads Jeppesen, Sigurd Madsen, Jenni Raitoharju, Maria Rasmussen, Alexandros Iosifidis, Ville Tirronen, Kristian Meissner, Moncef Gabbouj, and Toke Høye. Automatic image-based identification and biomass estimation of invertebrates. *Methods in Ecology and Evolution*, 11(8):922–931, 2020.

[2] Hu Cao, Guang Chen, Hengshuang Zhao, Dongsheng Jiang, Xiaopeng Zhang, and Qi Tian. SDPT: Semantic-aware dimension-pooling transformer for image segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):15934–15946, 2024.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

[4] Yongjie Chen, Hongmin Liu, Haoran Yin, and Bin Fan. Building vision transformers with hierarchy aware feature aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5908–5917, 2023.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[6] Arthur F.A. Fernandes, Eduardo M. Turra, Érika R. de Alvarenga, Tiago L. Passafaro, Fernando B. Lopes, Gabriel F.O. Alves, Vikas Singh, and Guilherme J.M. Rosa. Deep learning image segmentation for extraction of fish body measurements and prediction of body weight and carcass traits in Nile tilapia. *Computers and Electronics in Agriculture*, 170:105274, 2020.

[7] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant set input. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3744–3753. PMLR, 2019.

[8] Saebom Leem and Hyunseok Seo. Attention guided CAM: Visual explanations of vision transformer guided by self-attention. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

[9] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pages 12116–12128, 2021.

[10] Maria Sokolova, Manuel Cordova, Henk Nap, Aloysius van Helmond, Michiel Mans, Arjan Vroegop, Angelo Mencarelli, and Gert Kootstra. An integrated end-to-end deep neural network for automated detection of discarded fish species and their weight estimation. *ICES Journal of Marine Science*, 80(7):1911–1922, 2023.

[11] Kehao Wang, Yuhui Wang, Ruiqi Ren, Han Zou, and Zhichao Shao. Transformer-based feature aggregation and stitching network for crowd counting. *IEEE Access*, 11:124833–124844, 2023.

[12] Tianye Zhang, Yuqiao Yang, Yueyue Liu, Chenglei Liu, Ran Zhao, Daoliang Li, and Chen Shi. Fully automatic system for fish biomass estimation based on deep neural network. *Ecological Informatics*, 79:102399, 2024.

[13] Yuliang Zhao, Qijun Xiao, Jinhao Li, Kaixuan Tian, Le Yang, Peng Shan, Xiaoyong Lv, Lianjiang Li, and Zhikun Zhan. Review on image-based animals weight weighing. *Computers and Electronics in Agriculture*, 215:108456, 2023.

Équité des classifieurs en présence de contraintes entre les attributs

Martin C. Cooper[1], Imane Bousdira[2]

[1] IRIT, University of Toulouse
[2] IRIT, Toulouse INP

Firstname.LastName@irit.fr

Résumé

En apprentissage automatique, une définition admise de l'équité d'une décision prise par un classifieur est qu'elle ne doit pas dépendre d'attributs protégés, tels que le genre. Cependant, lorsque des contraintes existent entre les attributs, de telles dépendances peuvent être cachées par celles-ci. Pour pallier ce problème, nous proposons de qualifier une décision d'équitable si elle possède une explication équitable. Nous définissons cette explication comme un impliquant premier de la décision qui ne contient aucun attribut protégé, en tenant compte des contraintes.
Trois définitions possibles de l'équité d'un classifieur peuvent être distinguées : pour chacune de ses décisions, (1) toutes les explications sont équitables, (2) il existe au moins une explication équitable, ou (3) toute modification des attributs protégés n'affecte pas le résultat. Ce document présente un résumé de notre travail [8], dans lesquel nous avons identifié les relations entre ces différentes définitions et étudié la complexité computationnelle de la vérification de l'équité des classifieurs.

Mots-clés

Équité, contraintes, explication abductive, complexité.

Abstract

In Machine Learning, an accepted definition of fairness of a decision taken by a classifier is that it should not depend on protected features, such as gender. Unfortunately, when constraints exist between features, such dependencies can be obscured by the constraints. To avoid this problem, we propose that a decision be considered fair if it has a fair explanation. We define this explanation as a prime-implicant reason for the decision that does not contain any protected feature, while taking constraints into account.
Three possible definitions of fairness of a classifier are that for all its decisions (1) there are only fair explanations, (2) there is at least one fair explanation, or (3) changing protected features does not change the outcome. This document presents a summary of our work [8], in which we identified the relationships between these different definitions of fairness and studied the computational complexity of testing fairness of classifiers.

Keywords

Fairness, constraints, abductive explanation, complexity.

1 Introduction

L'attention portée à l'équité a considérablement augmenté ces dernières années, en raison de l'usage croissant de l'intelligence artificielle dans divers secteurs. Les préoccupations relatives à l'équité se sont intensifiées avec l'introduction des systèmes d'IA dans des domaines critiques, tels que les décisions de recrutement [20]. Par ailleurs, le principe d'équité a été largement mis en avant dans le cadre de l'IA de confiance [13, 17]. Étant donné l'importance de l'équité, ce concept a été l'objet de nombreuses études, donnant lieu à de multiples définitions [6, 19]. L'une des définitions les plus fréquemment citées décrit l'équité comme l'absence de toute forme de favoritisme ou de discrimination envers un individu ou un groupe, en fonction de caractéristiques innées ou acquises [16]. Des définitions plus précises sont proposées dans la littérature, que Verma et Rubin [22] analysent dans leur article de synthèse.
Il est raisonnable de considérer qu'un classifieur est équitable si aucune de ses décisions n'est influencée par des attributs protégés, tels que l'origine ethnique ou le genre. Cependant, en présence de contraintes entre les attributs, cette définition simple doit être affinée. Les contraintes peuvent provenir de différentes origines. Elles peuvent résulter de l'encodage (encodage one-hot dans l'exemple 1) ou de restrictions liées au monde réel. En effet, ignorer ces contraintes peut altérer l'évaluation de l'équité.

Exemple 1. *Considérons une fonction $\kappa(m, f, g)$ qui renvoie 1 si un employé est éligible à une prime en fonction des attributs booléennes : m (homme), f (femme), g (objectifs atteints). Les attributs m et f sont protégés. Supposons qu'il existe une contrainte unique : $m \equiv \neg f$. Pour cette tâche, un modèle possible est : $\kappa(m, f, g) \equiv (m \wedge g) \vee (f \wedge g)$. Sans tenir compte de la contrainte, ce modèle pourrait être considéré comme inéquitable puisque $\kappa(0, 0, 1) \neq \kappa(1, 0, 1)$ (deux instances qui ne diffèrent que par l'attibut m ont des résultats différents). Cependant, en considérant la constraint, $\kappa(m, f, g) \equiv g$, ainsi, κ est clairement équitable.*

Le reste du document présente un aperçu des principaux résultats détaillés dans l'article [8]. La contribution majeure réside dans la prise en compte des contraintes entre les attributs lors de l'évaluation de l'équité, en proposant une nouvelle définition de l'équité en termes d'explications.

2 Préliminaires

Soit $\mathcal{F}$ un ensemble de n attributs et $\mathcal{K}$ un ensemble de (au moins deux) classes. Nous partons du principe que l'ensemble des attributs $\mathcal{F}$ est divisé en un ensemble $\mathcal{P}$ d'attributs protégés (tels que l'origine ethnique ou le genre) et un ensemble $\mathcal{N}$ d'attributs non protégés. $\mathbb{F}$ désigne l'espace des attributs, qui est le produit cartésien des domaines des n attributs. Il se peut que certains vecteurs d'attributs de $\mathbb{F}$ ne soient pas possibles, en raison d'un ensemble de contraintes noté $\mathcal{C}$. Nous désignons par $\mathbb{F}[\mathcal{C}]$ l'ensemble des instances de $\mathbb{F}$ qui satisfont les contraintes. Nous supposons que $\kappa : \mathbb{F}[\mathcal{C}] \to \mathcal{K}$ est un classifieur. Pour simplifier la notation, nous associons à chaque attribut un nombre compris entre 1 et n. Pour $\mathcal{S} \subseteq \mathcal{F}$ et $x \in \mathbb{F}$, la notation $x_{\mathcal{S}}$ désigne l'affectation partielle $\{(i, x_i) : i \in \mathcal{S}\}$ aux attributs de $\mathcal{S}$.
Dans le cadre de l'intelligence artificielle explicable (XAI), une notion formelle d'explication d'une décision a émergé [1, 2, 4, 5, 15, 18, 21], souvent désignée par explication abductive AXp (ou raison suffisante).

Définition 1. *Une* AXp faible *d'une décision associée à un couple classifieur-instance $\langle \kappa, x \rangle$ est un sous-ensemble S d'attributs tel que $\forall y \in \mathbb{F}[\mathcal{C}]$, $y_S = x_S \to \kappa(y) = \kappa(x)$. Une* AXp *est une AXp faible minimal pour l'inclusion.*

Dans la définition d'une AXp, les contraintes $\mathcal{C}$ sont prises en compte en ne considérant que les instances y (qui coïncident avec x sur S) qui satisfont les contraintes [3, 9, 10, 23]. Selon le contexte et pour simplifier la présentation, il peut être utile de considérer l'affectation partielle x_S, plutôt que l'ensemble S, comme l'explication abductive.

3 Équité des décisions en présence de contraintes

Étant donné une décision $\kappa(x) = c$ dans un espace d'attributs *sans contrainte* $\mathbb{F}$, nous pouvons considérer une AXp comme un impliquant premier de la décision : il s'agit d'une raison suffisante pour la décision qui n'est pas subsumée par une autre raison suffisante. Dans un espace d'attibuts sans contrainte, A subsume B si et seulement si $A \subseteq B$. En présence de contraintes $\mathcal{C}$, la subsomption est plus subtile : lorsqu'on explique une décision $\kappa(x) = c$, *A subsume B* si et seulement si l'affectation partielle x_B implique l'affectation partielle x_A dans $\mathbb{F}[\mathcal{C}]$ (c'est-à-dire, $\forall y \in \mathbb{F}[\mathcal{C}], (y_B = x_B) \to (y_A = x_A)$). *$A$ subsume strictement B* si et seulement si A subsume B, mais B ne subsume pas A. Cela conduit à la définition 2 suivante [7]. Les définitions d'une décision équitable sont introduites ci-après.

Définition 2. *Une* PI-explication *(explication impliquant-premier) d'une décision $\kappa(x) = c$ dans un espace d'attributs contraint $\mathbb{F}[\mathcal{C}]$ est une AXp qui n'est pas strictement subsumé par une autre AXp.*

Définition 3. *Une PI-explication est* équitable *si elle ne comporte aucun attribut protégé. Nous disons qu'une décision est* existentiellement équitable *si elle possède une PI-explication équitable. Nous disons qu'une décision est* universellement équitable *si toutes ses PI-explications sont équitables. Une décision est* inéquitable *si elle ne possède aucune PI-explication équitable.*

Dans l'exemple 1, toutes les décisions prises par le classifieur sont universellement équitables, car elles ont toutes la même PI-explication unique $\{g\}$.

4 L'impact des contraintes sur l'équité des décisions

Dans un premier temps, nous examinons le cas où il existe des contraintes entre les attributs protégés $\mathcal{P}$ et les attributs non protégés $\mathcal{N}$. L'exemple 2 montre que l'ignorance des contraintes peut amener à considérer comme inéquitable une décision universellement équitable. Il convient de noter que le fait de ne pas tenir compte des contraintes peut également faire passer une décision inéquitable pour universellement équitable.

Exemple 2. *Considérons une fonction $\kappa(e, m)$ qui renvoie 1 si une personne est éligible à une formation gratuite, en fonction de deux attributs : 'en emploi' (e) et 'en congé de maternité' (m). Supposons que $\kappa(e, m) \equiv (\neg e) \wedge (\neg m)$. Sans tenir compte des contraintes, la seule PI-explication pour la décision $\kappa(0, 0) = 1$ est $\{e, m\}$. Si m est un attribut protégé, alors cette PI-explication est inéquitable et, par conséquent, la décision paraît inéquitable. Cependant, il existe une contrainte selon laquelle une personne ne peut être en congé maternité que si elle est employé, donc $(e, m) = (0, 1)$ est impossible. Ainsi, dans $\mathbb{F}[\mathcal{C}]$, $\kappa \equiv \neg e$ et la seule PI-explication pour $\kappa(0, 0) = 1$ est $\{e\}$, qui est équitable.*

La Figure 1 présente une synthèse de l'impact de la prise en compte ou l'ignorance des contraintes dans les cas où elles n'apparaissent qu'au sein de $\mathcal{P}$ ou qu'au sein de $\mathcal{N}$.

5 Équité des classifieurs en présence de contraintes

Dans cette partie, nous nous intéressons à l'équité des classifieurs, plutôt qu'à celle d'une décision individuelle. À cet égard, nous examinons une définition existante et en proposons de nouvelles.

5.1 FTU contrainte comme mesure d'équité

Rappelons la définition de FTU (Fairness Through Unawareness), qui est une règle classique permettant d'évaluer l'équité des classifieurs dans des espaces d'attributs non contraints [11, 14, 12] :

$$\forall x, y \in \mathbb{F},\ x_{\mathcal{N}} = y_{\mathcal{N}} \ \to\ \kappa(x) = \kappa(y)$$

Lorsque des contraintes existent, nous ne pouvons pas appliquer directement FTU comme indiqué ci-dessus, notamment parce que les valeurs de $\kappa(y)$ peuvent être considérées comme indéfinies pour $y \notin \mathbb{F}[\mathcal{C}]$. Comme le montre l'exemple 1, lorsque les contraintes sont ignorées, un classifieur équitable peut être considéré à tort comme inéquitable. Ainsi, la définition de FTU dans $\mathbb{F}[\mathcal{C}]$ est la suivante.

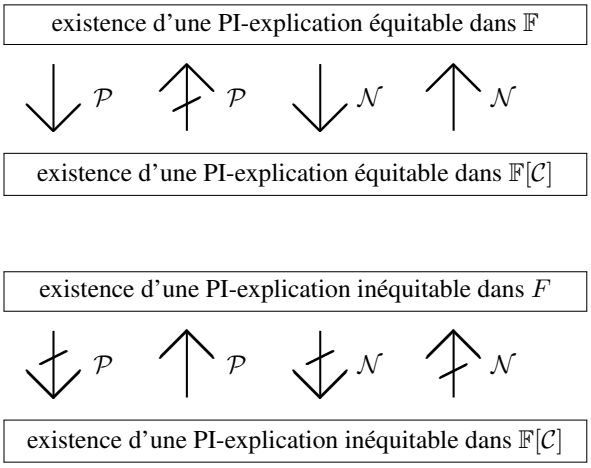


FIGURE 1 – La conservation (ou non) des PI-explications équitables/inéquitables de $\langle \kappa, x \rangle$ (pour $x \in \mathbb{F}[\mathcal{C}]$) lorsque les contraintes sont prises en compte (flèches vers le bas) ou ignorées (flèches vers le haut). $\mathcal{P}$ ou $\mathcal{N}$ signifie que les contraintes apparaissent uniquement au sein des attributs de $\mathcal{P}$ ou $\mathcal{N}$, respectivement.

Définition 4. *Un classifieur $\kappa : \mathbb{F}[\mathcal{C}] \to \mathcal{K}$ satisfait FTU dans $\mathbb{F}[\mathcal{C}]$* (FTU contrainte) *si*

$$\forall x, y \in \mathbb{F}[\mathcal{C}],\ x_{\mathcal{N}} = y_{\mathcal{N}} \ \rightarrow\ \kappa(x) = \kappa(y)$$

5.2 Mesures de l'équité basées sur les PI-explications

Conformément aux définitions d'une décision équitable données dans la Section 3, nous pouvons définir l'équité d'un classifieur, en se basant sur l'équité de ses décisions.

Définition 5. *Un classifieur $\kappa : \mathbb{F}[\mathcal{C}] \to \mathcal{K}$ satisfait*
- l'équité existentielle *si $\forall x \in \mathbb{F}[\mathcal{C}]$, $\langle \kappa, x \rangle$ a une PI-explication qui est équitable, et*
- l'équité universelle *si $\forall x \in \mathbb{F}[\mathcal{C}]$, toutes les PI-explications de $\langle \kappa, x \rangle$ sont équitables.*

6 Relations entre les définitions de l'équité des classifieurs

Nous résumons ci-dessous les principaux résultats issus de l'étude des relations entre les différentes définitions de l'équité des classifieurs.

- **FTU contrainte et équité existentielle**
  L'équité existentielle d'un classifieur κ implique que $\mathcal{N}$ est une AXp faible et donc que κ satisfait FTU dans $\mathbb{F}[\mathcal{C}]$. En revanche, FTU dans $\mathbb{F}[\mathcal{C}]$ n'implique pas l'équité existentielle.
- **Équité existentielle et équité universelle**
  Il est trivial qu'un classifieur satisfaisant l'équité universelle est existentiellement équitable. Cependant, l'équité existentielle n'implique pas l'équité universelle.

Nous introduisons par la suite des cas particuliers pour lesquels des définitions de l'équité s'avèrent équivalentes.

6.1 Absence de contraintes entre les attributs protégés et non protégés

Théorème 1. *Lorsqu'il n'existe aucune contrainte dans $\mathcal{C}$ entre les attributs protégés et non protégés, les notions suivantes d'équité d'un classifieur sont équivalentes :*

1. *équité universelle*
2. *équité existentielle*
3. *FTU contrainte (FTU dans $\mathbb{F}[\mathcal{C}]$)*

6.2 Contraintes lâches

Définition 6. *Soit un ensemble d'attributs $\mathcal{F}$ divisé en un ensemble d'attributs protégés $\mathcal{P}$ et un ensemble d'attributs non protégés $\mathcal{N}$. Un ensemble de contraintes $\mathcal{C}$ est* lâche *(par rapport à la partition $(\mathcal{P}, \mathcal{N})$ de l'ensemble des attributs) en $x \in \mathbb{F}[\mathcal{C}]$ si pour tout $p \in \mathcal{P}$, $x_{\{p\}}$ ne subsume pas strictement $x_{\mathcal{N}}$. L'ensemble de contraintes $\mathcal{C}$ est* lâche *si $\mathcal{C}$ est lâche en chaque $x \in \mathbb{F}[\mathcal{C}]$.*

Étant donné un ensemble de contraintes $\mathcal{C}$, la recherche d'affectations inéquitables qui susbsument strictement certaines affectations équitables peut être effectuée hors ligne en tant qu'étape préliminaire, indépendamment de tout classifieur, permettant de mieux comprendre les implications possibles liées à l'imposition de contraintes. Si de telles subsomptions n'existent pas, FTU contrainte garantit l'existence d'une PI-explication équitable.

Proposition 1. *Si l'ensemble de contraintes $\mathcal{C}$ est lâche, alors FTU dans $\mathbb{F}[\mathcal{C}]$ implique l'équité existentielle.*

6.3 Classifieurs détachés et contraintes

Définition 7. *Soit une instance $x \in \mathbb{F}[\mathcal{C}]$, la décision $\kappa(x) = c$ est* détachée *(par rapport aux contraintes $\mathcal{C}$) si l'ensemble $\mathcal{N}$ des attributs non protégés est une AXp faible de la décision $\kappa(x) = c$ qui n'est pas strictement subsumé par une AXp faible inéquitable. Un classifieur κ est* détaché *dans $\mathbb{F}[\mathcal{C}]$ si toutes ses décisions (pour $x \in \mathbb{F}[\mathcal{C}]$) sont détachées.*

Proposition 2. *Si un classifieur κ est détaché dans $\mathbb{F}[\mathcal{C}]$, alors κ est existentiellement équitable.*

7 Complexité computationnelle de la vérification de l'équité

Cette section présente les principaux résultats concernant la complexité de l'évaluation de l'équité. Nous supposons que κ est une fonction connue et que le calcul de $\kappa(z)$ pour $z \in \mathbb{F}[\mathcal{C}]$ peut être effectué en temps polynomial. Nous débutons par les tests d'équité existentielle et universelle pour une décision unique, à partir de laquelle nous déterminons la complexité de la vérification de l'équité d'un classifieur. Nous abordons ensuite la complexité du test de FTU contrainte qui est accessible aux solveurs SAT, avant de conclure par la complexité des tests pour évaluer si un ensemble de contraintes est lâche ou si un classifieur est détaché.

Proposition 3. *Le problème consistant à vérifier si une décision est existentiellement équitable appartient à* Σ_3^P. *Le problème consistant à vérifier si une décision est universellement équitable appartient à* Π_3^P.

Proposition 4. *Le problème consistant à vérifier l'équité existentielle d'un classifieur appartient à* Π_4^P. *Le problème consistant à vérifier l'équité universelle d'un classifieur appartient à* Π_3^P.

Proposition 5. *Le test de FTU contrainte pour un classifieur est coNP-complet.*

Nous pouvons énoncer le corollaire suivant du théorème 1, de la Proposition 1 et de la proposition 5.

Corollaire 1. *Lorsqu'il n'y a aucune contrainte dans* $\mathcal{C}$ *entre les attributs protégés et non protégés, le test de l'équité universelle d'un classifieur est coNP-complet. Lorsque l'ensemble de contraintes* $\mathcal{C}$ *est lâche, le test de l'équité existentielle d'un classifieur est également coNP-complet.*

Proposition 6. *Les problèmes consistant à vérifier si un ensemble de contraintes est lâche ou si un classifieur est détaché appartiennent à* Π_2^P.

8 Conclusion

L'équité est une exigence fondamentale pour les systèmes d'IA, dont l'influence croît dans des domaines critiques. L'évaluation de l'équité d'une décision particulière ou d'un classifieur dépend de la définition retenue de cette notion, étant donné la diversité des métriques existantes. Cette étude montre que l'ignorance des contraintes entre les attributs peut altérer de manière significative cette évaluation, notamment lorsqu'il existe des contraintes entre les attributs protégés et non protégés.

Les définitions proposées dans ce travail reposent sur l'existence d'explications ne contenant aucun attribut protégé. Nous considérons que cette approche est pertinente, car elle révèle les *raisons* sous-jacentes à chaque décision prise par le classifieur. Cependant, l'augmentation de la complexité computationnelle soulève certaines préoccupations. Pour y remédier, des conditions moins coûteuses à vérifier ont été identifiées.

Ce document présente un aperçu des principaux points abordés dans l'article [8], qui pose des bases théoriques novatrices sur l'équité formelle en présence de contraintes entre les attributs. L'approche théorique est enrichie par des perspectives pratiques, ouvrant ainsi plusieurs pistes pour la recherche future. L'article complet explore en profondeur ces éléments en fournissant des preuves, des exemples concrets et des explications détaillées.

Remerciements

Ce travail a été soutenu par le projet ForML ANR-23-CE25-0009.

Références

[1] Leila Amgoud and Jonathan Ben-Naim. Axiomatic foundations of explainability. In Luc De Raedt, editor, *IJCAI*, pages 636–642. ijcai.org, 2022. doi:10.24963/IJCAI.2022/90.

[2] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the computational intelligibility of boolean classifiers. In Meghyn Bienvenu, Gerhard Lakemeyer, and Esra Erdem, editors, *KR*, pages 74–86, 2021. doi:10.24963/KR.2021/8.

[3] Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. Deriving provably correct explanations for decision trees : The impact of domain theories. In *IJCAI*, pages 3688–3696. ijcai.org, 2024. URL : https://www.ijcai.org/proceedings/2024/408.

[4] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. Model interpretability through the lens of computational complexity. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *NeurIPS*, 2020.

[5] Shahaf Bassan, Guy Amir, and Guy Katz. Local vs. global interpretability : A computational complexity perspective. In *ICML*. OpenReview.net, 2024.

[6] Reuben Binns. Fairness in machine learning : Lessons from political philosophy. In Sorelle A. Friedler and Christo Wilson, editors, *Conference on Fairness, Accountability and Transparency, FAT*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159. PMLR, 2018.

[7] Martin C. Cooper and Leila Amgoud. Abductive explanations of classifiers under constraints : Complexity and properties. In Kobi Gal, Ann Nowé, Grzegorz J. Nalepa, Roy Fairstein, and Roxana Radulescu, editors, *ECAI*, pages 469–476. IOS Press, 2023.

[8] Martin C. Cooper and Imane Bousdira. Fairness of classifiers in the presence of constraints between features. In *CP 2026*, 2026. arXiv:2605.00592.

[9] Martin C. Cooper and João Marques-Silva. On the tractability of explaining decisions of classifiers. In Laurent D. Michel, editor, *CP*, volume 210 of *LIPIcs*, pages 21 :1–21 :18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPICS.CP.2021.21.

[10] Niku Gorji and Sasha Rubin. Sufficient reasons for classifier decisions in the presence of domain constraints. In *AAAI*, pages 5660–5667. AAAI Press, 2022. doi:10.1609/AAAI.V36I5.20507.

[11] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning : Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, 2016.

[12] Alexey Ignatiev, Martin C. Cooper, Mohamed Siala, Emmanuel Hebrard, and João Marques-Silva. Towards formal fairness in machine learning. In Helmut Simonis, editor, *CP*, volume 12333 of *LNCS*, pages 846–867. Springer, 2020.

[13] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. Trustworthy artificial intelligence : A review. *ACM Comput. Surv.*, 55(2), 2022.

[14] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, pages 4066–4076, 2017.

[15] João Marques-Silva and Alexey Ignatiev. Delivering trustworthy AI through formal XAI. In *AAAI*, pages 12342–12350. AAAI Press, 2022. doi:10.1609/AAAI.V36I11.21499.

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), 2021.

[17] High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019.

[18] Sebastian Ordyniak, Giacomo Paesani, Mateusz Rychlicki, and Stefan Szeider. Explaining decisions in ML models : A parameterized complexity analysis. In Pierre Marquis, Magdalena Ortiz, and Maurice Pagnucco, editors, *KR*, 2024. doi:10.24963/KR.2024/53.

[19] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. How do fairness definitions fare ? : Examining public attitudes towards algorithmic definitions of fairness. In Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor, editors, *AAAI/ACM Conference on AI, Ethics, and Society, AIES*, pages 99–106. ACM, 2019.

[20] Candice Schumann, Jeffrey S. Foster, Nicholas Mattei, and John P. Dickerson. We need fairness and explainability in algorithmic hiring. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith, editors, *AAMAS*, pages 1716–1720. International Foundation for Autonomous Agents and Multiagent Systems, 2020.

[21] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In Jérôme Lang, editor, *IJCAI*, pages 5103–5111. ijcai.org, 2018. doi:10.24963/IJCAI.2018/708.

[22] Sahil Verma and Julia Rubin. Fairness definitions explained. In Yuriy Brun, Brittany Johnson, and Alexandra Meliou, editors, *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE*, pages 1–7. ACM, 2018.

[23] Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey, Nina Narodytska, and João Marques-Silva. Eliminating the impossible, whatever remains must be true : On extracting and applying background knowledge in the context of formal explanations. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *AAAI*, pages 4123–4131. AAAI Press, 2023. doi:10.1609/AAAI.V37I4.25528.

SHGR: A Generalized Maximal Correlation Coefficient

Samuel Stocksieker[1], Denys Pommeret[1]

[1] CNRS, I2M, Aix-Marseille Université

samuel.stocksieker@univ-amu.fr

1 Résumé

1.1 Introduction de SHGR : une nouvelle mesure de corrélation non linéaire

La compréhension des dépendances entre variables constitue un enjeu central en apprentissage automatique, en statistique et en science des données, avec des applications essentielles telles que la sélection de variables, la réduction de dimension, l'évaluation de l'équité, l'inférence causale ou l'apprentissage multimodal. Les mesures classiques, en particulier les corrélations de Pearson et de Spearman, sont largement utilisées, mais se limitent aux relations linéaires/monotones et binaires, échouant souvent à saisir des dépendances plus complexes ou d'ordre supérieur. Pour pallier ces limitations, plusieurs mesures de dépendance généralisées ont été proposées. Parmi elles, la corrélation maximale de Hirschfeld-Gebelein-Rényi (HGR) se distingue comme un outil théoriquement fondé pour quantifier la dépendance non linéaire entre variables aléatoires, qu'elles soient univariées ou multivariées. Introduite par Hirschfeld ([2]), étendue par Gebelein ([1]) et formalisée par Rényi ([5]), la mesure HGR définit la corrélation comme la corrélation linéaire maximale entre des versions transformées des variables. Malgré son attrait théorique, l'estimation du coefficient HGR demeure un défi, en raison de la difficulté d'identifier les transformations optimales. D'autres méthodes présentent des inconvénients similaires, tels qu'une interprétabilité limitée, une complexité computationnelle élevée ou une difficulté d'extension à des configurations multivariées. De plus, la sensibilité des estimateurs HGR neuronaux aux valeurs aberrantes peut conduire à des comportements de sur-apprentissage et à des scores de corrélation artificiellement gonflés.

Pour remédier à ces limitations, nous proposons une extension naturelle du coefficient HGR. Notre approche s'inspire du coefficient de Spearman et des méthodes basées sur les copules, qui reposent sur les fonctions de répartition et sont intrinsèquement robustes aux valeurs extrêmes et aux relations non linéaires. Cette substitution offre deux avantages principaux : (i) elle améliore la robustesse aux valeurs extrêmes, ii) elle présente une invariance par transformation monotone et (iii) elle étend la portée de la dépendance mesurable pour inclure des relations monotones mais non linéaires. Il est important de noter qu'elle reste cohérente avec le principe HGR de maximisation de la corrélation entre les transformations non linéaires des variables, mais recadre la notion de dépendance en termes de monotonicité basée sur les rangs plutôt que de linéarité brute. Cela pourrait aider à réduire les efforts de transformation et à optimiser la calibration pour éviter de rechercher une corrélation linéaire, mais simplement une corrélation monotone. Nous rappelons que le coefficient de corrélation de Spearman entre deux échantillons appariés iid $u = (u_1, \cdots, u_n)$ et $v = (v_1, \cdots, v_n)$ issus de U et V, noté $\rho(u, v)$, est défini comme le coefficient de corrélation de Pearson basé sur les rangs :

$$\rho(u,v) = r\left(n\widehat{F}_U(u), n\widehat{F}_V(v)\right) = r\left(\widehat{F}_U(u), \widehat{F}_V(v)\right)$$
$$\text{avec } \widehat{F}_U(u) = (\widehat{F}_U(u_1), \cdots, \widehat{F}_U(u_n))$$

DEFINITION 1. (Coefficient Spearman-HGR (SHGR)). Soient U et V deux variables aléatoires continues appariées prenant leurs valeurs dans $\mathcal{U}$ et $\mathcal{V}$, respectivement. Soit $\mathcal{E}(\mathcal{U})$ (resp. $\mathcal{E}(\mathcal{V})$) l'ensemble des fonctions mesurables de $\mathcal{U}$ (resp. $\mathcal{V}$) vers $\mathbb{R}$. Le coefficient Spearman-HGR (SHGR) associé à (U, V) est défini par $SHGR(U, V) :=$

$$\max_{\substack{f_u\in\mathcal{E}(\mathcal{U}), f_v\in\mathcal{E}(\mathcal{V})\\ \mathbb{E}(f_u(U))=0, \mathbb{E}(f_v(V))=0\\ \mathbb{E}(f_u^2(U))=1, \mathbb{E}(f_v^2(V))=1}} r(F_{f_u(U)}(f_u(U)), F_{f_v(V)}(f_v(V))).$$

Il est important de noter que la transformation par copule préserve la dépendance entre les vecteurs originaux U et V (voir par exemple, [4]). En utilisant l'estimateur empirique $\widehat{F} : \widehat{F}_z(z_i) := \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{z_j \le z_i}$ de F, nous obtenons un estimateur du SHGR $\widehat{SHGR} =$:

$$\max_{\substack{f_u\in\mathcal{E}(\mathcal{U}), f_v\in\mathcal{E}(\mathcal{V})\\ \mathbb{E}(f_u(U))=0, \mathbb{E}(f_v(V))=0\\ \mathbb{E}(f_u^2(U))=1, \mathbb{E}(f_v^2(V))=1}} r(\widehat{F}_{f_u(U)}(f_u(U)), \widehat{F}_{f_v(V)}(f_v(V))),$$

et sa version neuronale $SHGR_\Theta(u, v) =$:

$$= \max_{f_{\theta_u}, f_{\theta_v}\in\Theta} \rho(f_{\theta_u}(u), f_{\theta_v}(v))$$
$$= \max_{f_{\theta_u}, f_{\theta_v}\in\Theta} r(\widehat{F}_{f_{\theta_u}(U)}(f_{\theta_u}(u)), \widehat{F}_{f_{\theta_v}(V)}(f_{\theta_v}(v)))$$

Notons que la transformation par rang est appliquée uniquement au calcul de la corrélation, et non aux variables d'entrée elles-mêmes. Cela préserve l'information des entrées originales tout en bénéficiant de la robustesse des objectifs basés sur les rangs. Typiquement, pour l'estimation bivariée, l'algorithme considère p variables d'entrée $u_1, \ldots, u_p$

et vise à optimiser les entrées correspondantes de la matrice de corrélation. La conception à encodeurs empilés permet des transformations marginales capturant des dépendances complexes et non linéaires entre les variables. La fonction objectif pour les estimations bivariées (corrélations par paires) est définie comme suit :

$$\mathcal{L}_{SHGR}(\boldsymbol{u}) := -\sum_{i,j=1\ i\neq j}^{p}\left[\rho^2\left(f_{\theta_{u_i}}(u_i), f_{\theta_{u_j}}(u_j)\right)^{1/2}\right]^{\alpha}$$

with $\alpha > 0$

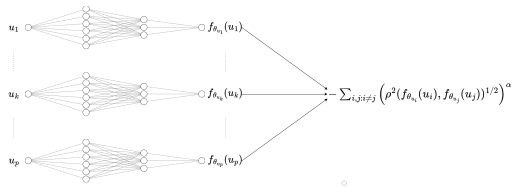
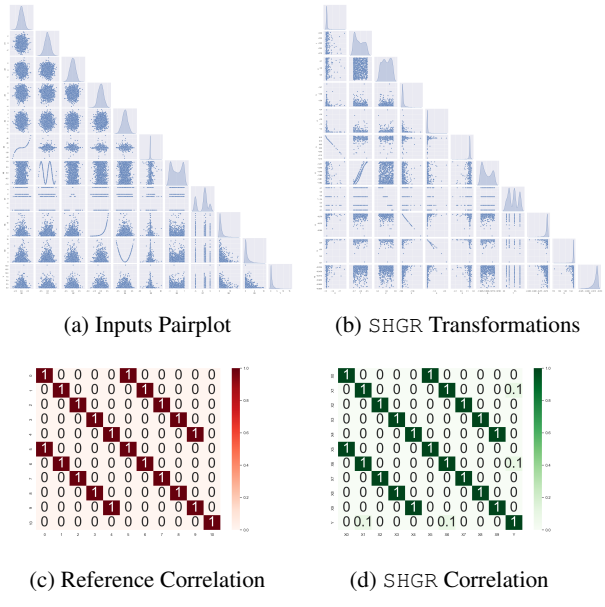


FIGURE 1 – Cross-Encoder Architecture

Ce nouveau coefficient estime des matrices de corrélation par paires, multivariées (p-vs-1 variables) et par groupes (p-vs-q variables). Cette extension s'appuie sur deux composantes clés : (i) une approximation neuronale conçue pour l'estimation simultanée de corrélations (Figure 1), et (ii) une formulation à base de copules qui améliore la robustesse et la stabilité en réduisant la sensibilité aux valeurs extrêmes.
La figure 2 présente une illustration de l'approche : l'algorithme transforme les données pour que les corrélations non linéaires deviennent linéaires.



(a) Inputs Pairplot (b) SHGR Transformations

(c) Reference Correlation (d) SHGR Correlation

FIGURE 2 – Illustration of SHGR on bivariate correlations

1.2 Protocole d'évaluation numérique

Nous proposons d'évaluer le SHGR et ses concurrents au moyen d'un protocole complet suivant l'analyse ci-dessous (définissant la *puissance multivariée des mesures de corrélation*) :

- Performance : capacité à capturer des corrélations non linéaires complexes sans bruit ;
- Robustesse au bruit : capacité à identifier des corrélations non linéaires complexes en présence de bruit ;
- Robustesse à l'hallucination : corrélation nulle en cas d'indépendance ;
- Robustesse aux valeurs extrêmes : corrélation nulle en cas d'indépendance, en présence de valeurs extrêmes ;
- *Puissance bivariée d'une mesure de dépendance* telle que proposée par [3] ;
- Temps de calcul : rapidité d'estimation des corrélations ;
- Analyse du test de significativité : possibilité d'effectuer un test de significativité sur le coefficient et capacité à rejeter l'hypothèse nulle (de corrélation nulle) en présence de bruit.

Nous appliquons le protocole de *puissance multivariée des mesures de corrélation* dans trois configurations : (i) corrélations bivariées (par paires) (1-vs-1), (ii) corrélations multivariées (p-vs-1), et (iii) corrélations complètes (par groupes) (p-vs-q). Pour (i), nous générons six variables gaussiennes indépendantes et cinq autres qui en dépendent de manière non linéaire (Figure 2a). Pour (ii), nous simulons 20 variables, incluant des dépendances non linéaires impliquant plus de deux variables. Pour (iii), nous générons deux jeux de données présentant diverses structures de corrélation globale. Nous estimons les corrélations non linéaires à l'aide de SHGR et les comparons aux méthodes alternatives suivantes (certaines n'étant pas disponibles dans les configurations multivariées et complètes) :

- Le coefficient de corrélation de Pearson (Pearson) ;
- Le coefficient de corrélation de Spearman (Spearman) ;
- Le coefficient de corrélation de Kendall (Kendall) ;
- Le Randomized Dependence Coefficient (RDC) ;
- Le Mutual Information Criterion (MIC) ;
- La Distance Correlation (dCor) ;
- L'Analyse en Corrélations Canoniques (CCA) ;
- L'Analyse en Corrélations Canoniques à noyau ; : cette approche n'a pas été intégrée car instable et trop coûteuse en temps de calcul
- L'Alternating Conditional Expectations (ACE) ;
- L'estimation HGR par noyau (HGRkde) ;
- L'estimation HGR par réseau neuronal (HGRnn) ;
- L'estimation HGR par treillis (HGRlat) ;
- L'estimation HGR par double noyau (dk) ;
- L'estimation HGR par noyau simple (sk) ;
- Le Normalized Hilbert-Schmidt Independence Criterion (HSIC) ;

— La corrélation de Hellinger : Cette approche n'a pas été intégrée car inefficace, limitée aux corrélations par paires et trop coûteuse en temps de calcul.

Nous entraînons notre modèle SHGR pendant 100, 200 et 500 epochs, ainsi qu'avec arrêt anticipé. Une analyse de sensibilité de l'architecture et des hyperparamètres de SHGR a été réalisée. Une architecture et une configuration d'hyperparamètres uniques ont ensuite été utilisées de manière cohérente pour l'ensemble des illustrations et expériences. Dans les analyses ci-dessous, nous générons $K = 10$ jeux de données synthétiques et comparons les corrélations estimées par chaque méthode aux corrélations vraies, connues, appelées corrélations de référence. Nous évaluons l'écart à ces corrélations de référence via : la matrice de corrélation par paires pour la corrélation bivariée, le vecteur de corrélations multiples pour la corrélation multivariée, et le coefficient de corrélation pour l'analyse de corrélation complète.

Les figures ci-dessous présentent les résultats obtenus dans le cas de corrélations bivariées. On remarque que le SHGR est plus performant et plus robuste que les compétiteurs. En effet, la mesure est la distance à la matrice de corrélation de référence : plus bas est a valeur meilleur est le coefficient.

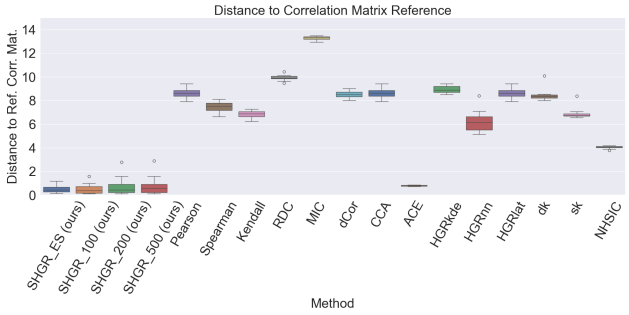


FIGURE 3 – Performance (Correlation bivariée)

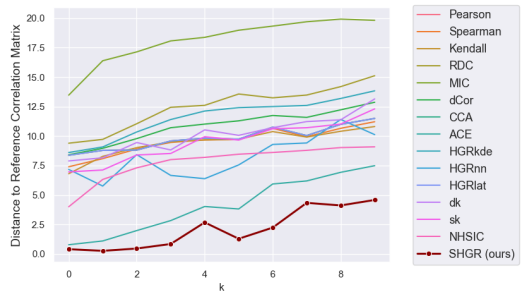


FIGURE 4 – Robustesse au bruit (Correlation bivariée)

Nos principales contributions peuvent être résumées comme suit :

— Nous revisitons la corrélation maximale HGR en proposant une extension basée sur Spearman : SHGR

— Nous présentons un premier estimateur de SHGR qui est i) ii) différentiable, iii) rapide, iv) efficace : détecte efficacement les liens non linéaires v) robuste au bruit :

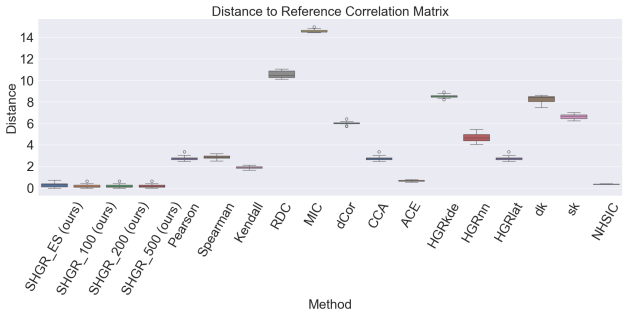


FIGURE 5 – Robustesse aux hallucinations (Correlation bivariée)

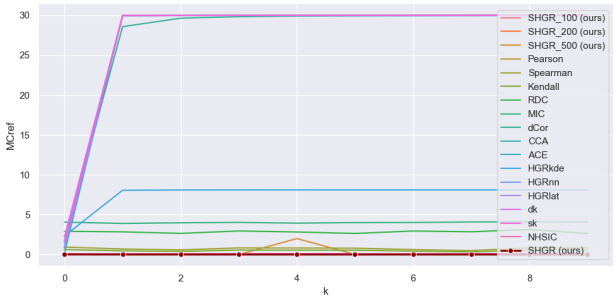


FIGURE 6 – Robustesse aux outliers (Correlation bivariée)

detecte les liens non linéaires en présence de bruit vi) robuste aux valeurs aberrantes : invariance du coefficient en présence de valeurs extrêmes et/ou abérantes et vii) robuste aux dépendances hallucinatoires : n'indique pas de corrélation signficatives lorsque les variales sont indépendantes. Contrairement à certaines méthodes antérieures, il récupère également les transformations de données et supporte les tests de significativité de la corrélation.

— Nous introduisons une architecture de cross-encodeur empilé spécifiquement conçue pour estimer simultanément (non itérative) plusieurs corrélations en contextes bivariés et multivariés.

— L'interprétabilité, généralement limitée pour les méthodes existantes, peut être partiellement améliorée via les transformations apprises accessibles dans SHGR, ouvrant des pistes d'analyse visuelle.

— Nous déduisons un protocole d'évaluation complet, la *Puissance multivariée de la mesure de corrélation*, pour évaluer les estimateurs de corrélation maximale en termes de performance, robustesse au bruit, hallucinations (respectant l'Axiome 4 de Rényi), valeurs extrêmes, ainsi que l'estimation de corrélations bivariées, multivariées et complètes, tests de significativité et efficacité computationnelle.

— Nous validons SHGR sur des ensembles de données tabulaires synthétiques et réels. Une évaluation sur des applications réelles de sélection de vairables est également réalisée, sur 9 datasets. Nous démontrons que SHGR surpasse les méthodes de l'état de l'art exis-

tantes en termes de performance et de robustesse. L'article original est disponible à l'adresse : https://neurips.cc/virtual/2025/loc/san-diego/poster/117128. SHGR constitue une base prometteuse pour de futurs travaux, notamment en haute dimension, en architectures à grande échelle, en apprentissage multimodal ou dans des contextes d'équité algorithmique.

Mots-clés

Corrélation non linéaire

Références

[1] Hans Gebelein. Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *Zammzeitschrift Fur Angewandte Mathematik Und Mechanik*, 21 :364–379, 1941.

[2] H. O. Hirschfeld. A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4) :520–524, 1935.

[3] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. *Advances in neural information processing systems*, 26, 2013.

[4] Roger B. Nelsen. *An introduction to copulas.* Springer, New York, 2006.

[5] Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10 :441–451, 1959.

SESSION 9 : Ontologies et logiques de description présidée parNathalie Aussenac

Induction de concepts pour la Qualité 4.0 explicable

Léa Charbonnier[1], Franco Giustozzi[2], Julien Saunier[1], Cecilia Zanni-Merk[1]

[1] INSA Rouen Normandie, Univ Rouen Normandie, Université Le Havre Normandie, Normandie Univ, LITIS UR 4108, F-76000 Rouen, France
[2] INSA Strasbourg, University of Strasbourg, ICube laboratory, CNRS (UMR 7357), 67000 Strasbourg, France

{lea.charbonnier, julien.saunier, cecilia.zanni-merk}@insa-rouen.fr, franco.giustozzi@insa-strasbourg.fr

Résumé

L'article présente une méthode permettant de diagnostiquer les causes des problèmes de qualité des produits en utilisant l'induction de concepts. Une nouvelle métrique d'évaluation, combinant la couverture des instances, la précision de la classification et la pertinence hiérarchique, est proposée pour identifier les concepts les plus informatifs. L'approche est validée à l'aide d'un ensemble de données synthétiques de maintenance prédictive. Les résultats montrent une identification précise des facteurs liés aux produits défectueux.

Mots-clés

Induction de concepts, Assurance qualité et Industrie 4.0 (Qualité 4.0), Ontologie.

Abstract

The article presents a method for diagnosing the causes of product quality issues using concept induction. A new evaluation metric that combines instance coverage, classification accuracy, and hierarchical relevance is proposed to identify the most informative concepts. The approach is validated using a synthetic predictive maintenance dataset. The results show accurate identification of factors related to defective products.

Keywords

Concept Induction, Quality Assurance and Industry 4.0 (Quality 4.0), Ontology.

Contexte

L'un des principaux défis pour l'industrie consiste à détecter rapidement les anomalies et à en identifier les causes racines afin de minimiser les répercussions économiques et écologiques [7]. Dans ce contexte, la Qualité 4.0 intègre des systèmes intelligents afin d'améliorer la gestion de la qualité. Il s'agit d'une extension de l'Industrie 4.0 qui utilise des technologies intelligentes dans les processus d'assurance qualité, c'est-à-dire les procédures visant à garantir que les produits répondent à des normes prédéfinies [1]. Les méthodes basées sur les données permettent d'identifier efficacement certaines causes racines, mais elles nécessitent généralement de grands ensembles de données étiquetées [9] et reposent sur des modèles complexes dont les processus décisionnels sont difficiles à interpréter [4]. Ces approches privilégient souvent la performance prédictive au détriment de l'explicabilité, ce qui peut limiter leur adoption dans des contextes industriels où il est essentiel de comprendre le raisonnement à l'origine d'un diagnostic. L'induction de concepts constitue une alternative interprétable, permettant d'extraire automatiquement des concepts en logique de description [6, 8]. Formellement, étant donné une base de connaissances $\mathcal{O} = (\mathcal{T}, \mathcal{A})$, composée d'une TBox $\mathcal{T}$ et d'une ABox $\mathcal{A}$, ainsi que deux ensembles disjoints d'individus P et N représentant respectivement des exemples positifs et négatifs, l'objectif est d'apprendre une expression de classe C telle que $\mathcal{O} \models C(a)$ pour tout $a \in P$ et $\mathcal{O} \not\models C(b)$ pour tout $b \in N$.

L'article original est paru à RuleML+RR 2025 [2]. Il propose une approche fondée sur l'induction de concepts et une représentation ontologique du domaine, combinées à une stratégie d'évaluation multicritère permettant d'identifier des concepts interprétables expliquant les causes potentielles des défauts de qualité.

Approche proposée : RCIQ4.0

L'article propose une nouvelle approche pour identifier des problèmes de qualité dans les processus de fabrication. Cette approche, illustrée par la figure 1, utilise l'induction de concepts pour découvrir et expliquer les causes de ces problèmes. Un élément clé de cette approche est une ontologie permettant de représenter les produits défectueux (exemples positifs), les produits conformes (exemples négatifs) et les situations machine. Celle-ci s'appuie sur l'ontologie contextuelle *Context Ontology* [3], qui fournit une représentation formelle du domaine de la fabrication. Le module *Concept Induction* utilise l'algorithme ECII [6] pour générer des expressions conceptuelles décrivant les produits défectueux.

ECII est un algorithme d'induction de concepts qui sé-

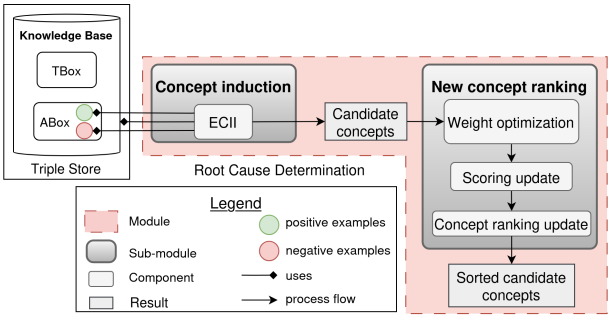


FIGURE 1 – Architecture de RCIQ4.0

lectionne un ensemble fini $\{C_1, \ldots, C_n\}$ d'expressions de classes complexes sur $\mathcal{L}$ ainsi que des expressions de classes complexes supplémentaires, ce qui donne $\mathcal{O}' = \mathcal{O} \cup \{A_k \equiv C_k | k = 1, \ldots, n\}$ où A_i sont des classes atomiques n'apparaissant pas dans $\mathcal{O}$. Ensuite, un raisonneur est utilisé pour calculer l'appartenance de tous les individus issus des exemples à toutes les classes atomiques de $\mathcal{O}'$. Enfin, des expressions de classe candidates sont générées à l'aide de constructeurs de la logique de description (conjonction, disjonction, négation et restrictions existentielles). Cette stratégie minimise le nombre d'appels au raisonneur, en pré-calculant les appartenances aux classes candidates, ce qui permet une génération plus efficace des concepts et une amélioration des performances par rapport aux approches basées sur un raffinement itératif. Cependant, cette approche a pour contrepartie de restreindre l'espace de recherche des expressions potentielles. En effet, la grammaire ne décrit pas d'expressions $\mathcal{ALC}$ arbitraires, mais une forme normale restreinte correspondant à la structure des concepts candidats générés par ECII, typiquement des combinaisons de classes atomiques et de restrictions existentielles, par exemple $\exists r.(A \sqcup B)$ ou $A \sqcap \exists r.B$. ECII est donc choisi pour son efficacité dans la génération d'expressions de classes complexes en logique descriptive, en particulier dans les ontologies riches en axiomes. Cependant, comme la méthode peut produire des centaines de concepts candidats, un processus de classement efficace devient essentiel pour déterminer quelles expressions méritent d'être examinées plus en détail.

Dans sa version originale, ECII classe les concepts uniquement selon leur couverture, ce qui tend à privilégier des concepts trop généraux. Afin d'obtenir des concepts plus discriminants, nous proposons d'évaluer chaque concept selon trois critères complémentaires : couverture, précision et hiérarchie. La couverture normalisée (Eq. 1) représente les exemples positifs correctement classés, c'est-à-dire les produits qui ont échoué au contrôle qualité et qui sont correctement décrits par le concept. La précision (Eq. 2) correspond à la proportion de vrais positifs déduits par le concept. La hiérarchie (Eq. 3) correspond à la profondeur ontologique du concept parmi tous les concepts de classe candidats. L'objectif est d'évaluer si le concept cible les produits défectueux (vrais positifs) et évite les bons (faux positifs), ainsi que son ratio de généralisation et de spécificité.

$$\text{n_cov} = \frac{\frac{\text{Nb de produits défectueux identifiés}}{\text{Nb total de produits défectueux}}}{\text{Couverture maximale}} \quad (1)$$

$$\text{precision} = \frac{\text{Nb de produits défectueux identifiés}}{\text{Nb total de produits identifiés}} \quad (2)$$

$$\text{hierarchy} = \frac{\text{Profondeur du concept}}{\text{Profondeur maximale}} \quad (3)$$

Les concepts sont ensuite classés en combinant les trois critères d'évaluation précédemment définis. Leur complémentarité permet un classement plus pertinent. Comme chaque critère est normalisé dans une plage comprise entre 0 et 1, ils peuvent être efficacement agrégés à l'aide d'une somme pondérée (équation 4).

$$\text{score} = \alpha \times \text{n_cov} + \beta \times \text{precision} + \gamma \times \text{hierarchy}$$
$$\text{avec } \alpha, \beta, \gamma \in [0, 1] \text{ et } \alpha + \beta + \gamma = 1 \quad (4)$$

Afin d'identifier la combinaison optimale des poids α, β et γ, la fonction de notation définie dans l'équation 4 est optimisée sous certaines contraintes.
La méthodologie de classement proposée est appliquée à un cas d'utilisation afin d'évaluer son efficacité dans l'identification et l'interprétation des causes racines.

Étude de cas et résultats

Pour évaluer l'approche RCIQ4.0, le jeu de données Predictive Maintenance [5] est utilisé. Il fournit des informations sur les produits (ID et type), les valeurs des capteurs et les données sur les pannes des machines. En l'absence d'observations sur le contrôle qualité, des étiquettes de contrôle qualité synthétiques basées sur les valeurs et les descriptions de l'ensemble de données sont générées. Elles permettent de diviser les produits selon deux ensembles (défectueux et en bon état) qui sont utilisés pour l'induction de concept.
À partir des concepts générés par ECII et de leurs métriques associées, les poids α, β et γ sont déterminés en résolvant un problème d'optimisation. Les valeurs obtenues sont $\alpha = \beta = \gamma = 0.333$, indiquant que les trois critères contribuent de manière équilibrée au score final. Ces pondérations sont spécifiques au jeu de données et aux concepts générés et doivent être recalculées si ceux-ci sont modifiés. Les trois concepts les mieux classés selon leur score pondéré sont présentés dans le tableau 1. Les valeurs de précision élevées (Prec. = 1.0) indiquent que les concepts identifient uniquement des produits défectueux, tandis que les scores de hiérarchie élevés traduisent leur spécificité. Chaque concept associe les produits défectueux à une ou deux situations machine, ce qui facilite leur interprétation. La propriété d'objet `involvedInSituation`, qui relie un produit à une situation machine, permet ainsi d'identifier les conditions anormales associées aux défauts. Par

TABLE 1 – Concepts et leurs métriques associées : couverture, hiérarchie, précision et score (Eq. 1, 3, 2 et 4). La relation `p` représente la propriété `involvedInSit.`

Concepts	NCov.	Hier.	Prec.	Score
∃p.((Sit-PWF) ⊔ (Sit-HDF))	0.742	0.75	1.000	**0.831**
∃p.(Sit-HDF)	0.452	**1.00**	1.000	0.817
∃p.((Sit-OSF) ⊔ (Sit-HDF))	0.645	0.75	1.000	0.798

exemple, le premier concept relie les produits défectueux aux situations `Power Failure (PWF)` ou `Heat Dissipation Failure (HDF)`, suggérant que ces défaillances peuvent être à l'origine du problème de qualité. Ainsi, la combinaison des métriques de couverture, de précision et de hiérarchie permet de sélectionner des concepts à la fois précis et interprétables, fournissant des explications potentielles aux défauts observés.

Discussion

Contrairement aux approches basées uniquement sur la couverture, la méthode proposée permet de privilégier des concepts plus spécifiques et interprétables, caractérisant les produits défectueux, mais elle présente certaines limites. En particulier, la qualité des concepts induits dépend fortement de la représentation ontologique et des données disponibles. De plus, la restriction de l'espace de recherche dans ECII peut limiter l'expressivité des concepts générés. Enfin, le choix des poids dans la fonction de score, bien qu'optimisé, reste dépendant du jeu de données considéré.

Conclusion

Cet article présente RCIQ4.0, une approche visant à identifier les causes potentielles de défauts de qualité au sein d'un processus de fabrication. La méthode repose sur l'induction de concepts afin d'extraire des concepts discriminants caractérisant les produits défectueux. Une nouvelle stratégie d'évaluation combinant couverture, précision et profondeur hiérarchique est proposée afin d'améliorer le classement des concepts. Cette mesure composite facilite l'identification de concepts explicatifs permettant de mieux comprendre l'origine des défauts de qualité.
L'approche proposée est générique et peut être appliquée à d'autres tâches d'analyse explicable basées sur des ontologies, dès lors que des exemples positifs et négatifs sont disponibles.

Remerciements

Ce travail a été financé par l'Agence nationale de la recherche française (numéro de subvention ANR-22-CE92-0007).

Références

[1] Adriana Ventura Carvalho, Daisy Valle Enrique, Amal Chouchene, and Fernando Charrua-Santos. Quality 4.0 : An Overview. *Procedia Computer Science*, 181 :341–346, 2021.

[2] Léa Charbonnier, Franco Giustozzi, Julien Saunier, and Cecilia Zanni-Merk. A Novel Concept Induction Approach for Explainable Quality 4.0. In *Rules and Reasoning*, pages 36–53, Cham, 2025. Springer Nature Switzerland.

[3] Franco Giustozzi, Julien Saunier, and Cecilia Zanni-Merk. Context Modeling for Industry 4.0 : an Ontology-Based Proposal. *Procedia Computer Science*, 126 :675–684, 2018.

[4] Hao Huang, Tapan Shah, John Karigiannis, and Scott Evans. Deep Root Cause Analysis : Unveiling Anomalies and Enhancing Fault Detection in Industrial Time Series. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, June 2024.

[5] Stephan Matzka. Explainable artificial intelligence for predictive maintenance applications. In *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*, pages 69–74, 2020.

[6] Md Kamruzzaman Sarker and Pascal Hitzler. Efficient concept induction for description logics. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, volume 33, pages 3036–3043. AAAI Press, January 2019.

[7] Md. Abdus Shabur. A comprehensive review on the impact of Industry 4.0 on the development of a sustainable environment. *Discover Sustainability*, 5(1) :97, May 2024.

[8] Cara Leigh Widmer and et al. Towards human-compatible XAI. *Journal of Web Semantics*, 79 :100807, December 2023.

[9] Chen Zhang, Di Hu, and Tao Yang. Research of artificial intelligence operations for wind turbines considering anomaly detection, root cause analysis, and incremental training. *Reliability Engineering & System Safety*, 241 :109634, January 2024.

Mesures de responsabilité à faible complexité pour requêtes en présence d'ontologies

Meghyn Bienvenu, Diego Figueira, Pierre Lafourcade

Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR5800, F-33400 Talence, France

{meghyn.bienvenu, diego.figueira, pierre.lafourcade}@u-bordeaux.fr

Résumé

Des travaux récents visant à expliquer des réponses à des requêtes ont mis au jour différentes mesures de responsabilité, qui attribuent un score à chaque fait d'une base de données, afin quantifier sa contribution à une réponse donnée. Dans cet article, nous étudions la complexité du calcul de telles mesures en présence d'ontologies. Nous nous concentrons sur une famille de mesures récemment introduites à partir de la valeur de Shapley, et qui s'expriment comme des sommes pondérées de supports minimaux (abrégé WSMS en anglais), dont nous explorons à la fois la complexité combinée et celle des données.

Mots-clés

Requêtes en présence d'ontologies, mesures de responsabilité, valeur de Shapley, analyse de complexité.

Abstract

Recent work on quantitative approaches to explaining query answers employs responsibility measures to assign scores to facts in order to quantify their respective contributions to obtaining a given answer. In this paper, we study the complexity of computing such responsibility scores in the setting of ontology-mediated query answering, focusing on a very recently introduced family of Shapley-value-based responsibility measures defined in terms of weighted sums of minimal supports (WSMS), exploring both their data and combined complexities for various ontology languages.

Keywords

Ontology-mediated query answering, responsibility measures, Shapley value, complexity analysis

☞ Cet article est un résumé de [3].

1 Introduction

Parmi les méthodes étudiées pour expliquer les réponses aux requêtes, les mesures quantitatives de responsabilité attribuent un score à chaque fait d'une base de données quantifiant sa contribution à la réponse d'une requête. Nous ne considèrerons ici que des requêtes booléennes et croissantes ;[1] il s'agira donc d'expliquer pourquoi la requête est satisfaite ($\mathcal{D} \models q$). Deux mesures notables ont été définies à partir de la valeur de Shapley Sh [5], méthode de distribution des richesses dans un jeu coopératif définie dans les années 50 : la valeur de Shapley dite *drastique* [4], et celle dite *MS*, plus récente [2]. Elles sont obtenues en modélisant la requête q et la base de données $\mathcal{D}$ par deux jeux différents : d'après le jeu $\xi^{\mathsf{dr}}_{q,\mathcal{D}}$ de la valeur dite drastique, une sous-base $S \subseteq \mathcal{D}$ reçoit 1 si $S \models q$ (et 0 sinon), tandis que pour le jeu $\xi^{\mathsf{ms}}_{q,\mathcal{D}}$ sous-jacent à la valeur MS, une sous-base S reçoit le nombre de *supports minimaux* (i.e. sous-ensembles minimaux M de faits tels que $M \models q$) présents dans S.

Dans le présent article, nous nous intéressons au calcul de la valeur de Shapley MS sur des données enrichies de connaissances au moyen d'une ontologie, à la suite de [1] qui avait fait de même pour la valeur de Shapley drastique. En effet, les résultats obtenus dans le cadre des bases de données (sans ontologie) ont montré que valeur de Shapley MS est généralement plus facile à calculer que la valeur drastique. De fait, $\mathrm{Sh}_{\xi^{\mathsf{ms}}_{q,\mathcal{D}}}(f)$ s'exprime plus simplement comme la somme, sur tous les supports minimaux M contenant f, de $\frac{1}{|M|}$. En outre, la valeur de Shapley MS se généralise à toute une famille de mesures (WSMS pour *weighted sums of minimal supports*) qui s'expriment facilement sous forme close, tout en restant chacune la valeurs de Shapley d'un certain jeu coopératif la caractérisant.

De plus, la valeur de Shapley MS (comme l'ensemble des mesures WSMS) est semblable à la valeur de Shapley drastique sur le plan conceptuel puisque toutes satisfont tous les desiderata suivants introduits dans [2, §4.1] et illustrés dans l'Exemple 1 : *(Sym-db)* deux faits indistinguables pour q doivent avoir la même valeur ; *(Null-db)* un fait, dit *pertinent*, qui appartient à un support minimal doit avoir une valeur strictement positive, et un fait non pertinent doit avoir une valeur nulle ; toutes choses égales par ailleurs, *(MS1)* un fait apparaissant dans des supports minimaux plus petits a une plus grande valeur, et *(MS2)* un fait apparaissant dans plus de supports minimaux a une plus grande valeur.

Exemple 1. *Considérons la base de connaissances $(\mathcal{D}, \mathcal{O})$ présentée en figure 1, ainsi que la requête booléenne $q :=$ $\mathsf{AuPoisson}(soleCancalaise)$. Il existe trois supports mini-*

1. Une requête est *booléenne* si elle est sans variables libres (une question fermée : «vrai» ou «faux»), et *croissante* si l'ajout de nouveaux faits préserve les réponses (ou la satisfaction). Des réponses aux requêtes avec variables libres sont traitées par réduction aux requêtes booléennes.

$\exists \mathsf{Ingr}.\mathsf{AuPoisson} \sqsubseteq \mathsf{AuPoisson}$ $\quad \mathsf{Garniture} \sqsubseteq \mathsf{Ingr}$

$\mathsf{FruitMer} \sqsubseteq \mathsf{AuPoisson}$ $\quad \mathsf{Poisson} \sqsubseteq \mathsf{AuPoisson}$

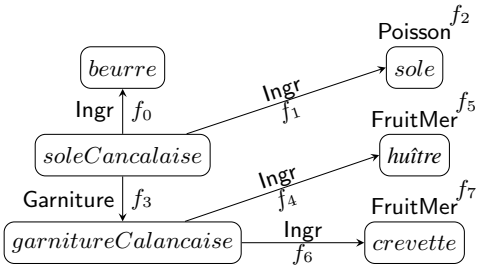


FIGURE 1 – Base de connaissances en logique de description comprenant des données $\mathcal{D}$ (bas) et une ontologie $\mathcal{O}$ (haut). Les flèches étiquetées indiquent des faits binaires et les étiquettes des nœuds (ex. Poisson) des faits unaires.

$\star$	f_0	f_1, f_2	f_3	f_4, f_5, f_6, f_7
dr	0	$\frac{1224}{5040}$	$\frac{1056}{5040}$	$\frac{384}{5040}$
ms	0	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{3}$

TABLE 1 – Valeurs de $\mathrm{Sh}_{\xi^{\star}_{q,\mathcal{D}}}(_)$ pour la figure 1. Les faits équivalents par (Sym-db) sont regroupées.

maux pour $Q := (\mathcal{O}, q)$ dans $\mathcal{D}$: $\{f_1, f_2\}$, $\{f_3, f_4, f_5\}$ et $\{f_3, f_6, f_7\}$. Les propriétés décrites ci-dessus s'appliquent comme suit : (Sym-db) donne $\phi_{Q,\mathcal{D}}(f_1) = \phi_{Q,\mathcal{D}}(f_2)$ (car f_1 et f_2 apparaissent dans les mêmes supports minimaux); (Null-db) donne $\phi_{Q,\mathcal{D}}(f_0) = 0$ (car f_0 n'est pas pertinent); (MS1) donne $\phi_{Q,\mathcal{D}}(f_1) > \phi_{Q,\mathcal{D}}(f_4)$ (car le support minimal où apparaît f_1 est plus petit); (MS2) donne $\phi_{Q,\mathcal{D}}(f_3) > \phi_{Q,\mathcal{D}}(f_4)$ (car f_3 apparaît dans deux supports minimaux contre un seul pour f_4). △

2 Complexité

Complexité des données Etudions la complexité du problème $\mathsf{SVC}^{\mathsf{ms}}_{Q}$ de calculer $\mathrm{Sh}_{\xi^{\mathsf{ms}}_{Q,\mathcal{D}}}(f)$ pour un base de données $\mathcal{D}$ et un fait f en entrée, avec une requête en présence d'ontologie $Q = (\mathcal{O}, q)$ fixée. Comme nous l'avons vu, cette tâche revient à calculer une somme pondérée de supports minimaux, et donc à dénombrer les supports minimaux groupés par taille. Toute requête en présence d'ontologie pouvant se réécrire en *union de requêtes conjonctives* (UCQ) n'admettant que des supports minimaux de taille bornée (par la taille l de la réécriture), il est possible de les énumérer naïvement en temps polynomial ($\mathcal{D}^l$) pour les dénombrer et par suite en déduire $\mathsf{SVC}^{\mathsf{ms}}_{Q}$:

Theorem 2. $\mathsf{SVC}^{ms}_{Q} \in \mathsf{FP}$ *pour toute requête en présence d'ontologie Q admettant une réécriture en UCQ.*[2] *C'est en particulier le cas de tout $Q \in (\mathsf{DL\text{-}Lite}_{\mathcal{R}}, \mathsf{UCQ})$.*

2. Rappelons que toutes les requêtes considérées ici sont croissantes.

Pour le calcul pratique, cette énumération naïve peut, en réalité, être nettement améliorée puisque l'évaluation de $\mathsf{SVC}^{\mathsf{ms}}_{q}$ pour toute UCQ q peut s'effectuer par l'évaluation en parallèle d'un ensemble de requêtes SQL dérivées de q.

Si l'on s'écarte des requêtes admettant une réécriture en UCQ, en revanche, l'on trouve très vite des requêtes demandant intuitivement «existe-t-il un chemin entre a et b ?», i.e. des requêtes exprimant la notion d'accessibilité dans un graphe. Or les supports minimaux pour de telles requêtes sont les chemins simples, qui sont #P-difficiles à compter.

Theorem 3. *Soit $\mathcal{L}$ une logique de description pouvant exprimer l'axiome $\exists r.A \sqsubseteq A$. Alors il existe une requête $(\mathcal{O}, q)$ avec $\mathcal{O} \in \mathcal{L}$ et q atomique tq SVC^{ms}_{Q} est #P-difficile.*

Complexité combinée Si nous avons établi que $\mathsf{SVC}^{\mathsf{ms}}_{Q} \in \mathsf{FP}$ pour tout Q se réécrivant en UCQ, la complexité combinée (qui ne suppose donc plus que $|Q|$ est constante) devient #P-difficile à la moindre conjonction dans les axiomes :

Proposition 4. SVC^{ms}_{Q} *est #P-diff. pour la logique de description $\mathcal{L}_{\sqcap}$ autorisant les axiomes de la forme $A \sqcap B \sqsubseteq C$.*

Si, en revanche, l'on se restreint à des logiques comme $\mathsf{DL\text{-}Lite}_{\mathcal{R}}$ pour lesquelles les supports minimaux de requêtes atomiques sont toujours des singletons, il est possible de recouvrer certains des résultats positifs connus pour les requêtes sans onotologie. Plus précisément, l'on sait que $\mathsf{SVC}^{\mathsf{ms}}_{\mathcal{C}} \in \mathsf{FP}$ pour toute classe $\mathcal{C}$ de *requête conjonctive* (CQ) sans auto-jointure (i.e. sans atomes distincts du même prédicat) et de largeur d'arbre bornée. La notion d'auto-jointure n'est pas pertinente en présence d'ontologie, parce que des axiomes $\{A \sqsubseteq B, B \sqsubseteq A\}$ peuvent confondre les prédicats A et B, mais l'on peut capturer son esprit par la notion d'*interaction*, qui intervient dès qu'un fait unique peut satisfaire deux atomes distincts de la requête.[3]

Theorem 5. *Pour toute classe $\mathcal{C} \subseteq (\mathsf{DL\text{-}Lite}_{\mathcal{R}}, \mathsf{CQ})$ telle que tout $Q \in \mathcal{C}$ est sans interaction et que l'ensemble $\{q \mid (\mathcal{O}, q) \in \mathcal{C}\}$ ait une largeur d'arbre bornée, $\mathsf{SVC}^{ms}_{\mathcal{C}} \in \mathsf{FP}$.*

Références

[1] Meghyn Bienvenu, Diego Figueira, and Pierre Lafourcade. Shapley value computation in ontology-mediated query answering. In *KR*, 2024.

[2] Meghyn Bienvenu, Diego Figueira, and Pierre Lafourcade. Shapley revisited : Tractable responsibility measures for query answers. In *PODS*, 2025.

[3] Meghyn Bienvenu, Diego Figueira, and Pierre Lafourcade. Tractable responsibility measures for ontology-mediated query answering. In *KR*, 2025.

[4] Ester Livshits, Leopoldo Bertossi, Benny Kimelfeld, and Moshe Sebag. The Shapley value of tuples in query answering. *LMCS*, Volume 17, Issue 3 :6942, 2021.

[5] Lloyd Stowell Shapley. A value for n-person games. *Contributions to the Theory of Games*, Volume II :307–318, 1953.

3. La véritable notion d'interaction est plus subtile : il est notamment possible d'avoir une interaction avec un seul atome, mais l'idée reste celle d'un fait satisfaisant l'atome de deux « manières différentes ».

SESSION 10 : IA générative, LLM et prompts présidée par Safa Yahi

Revue des travaux récents sur l'usage des LLMs en cybersécurité offensive

Hamza Bendali[1], Oussema Ben Henni[1], Youssef Douzi[1], Abokor Mahmmad Mousse[1], Frédéric Flouvat [2], Safa Yahi[2]

[1]Université d'Aix-Marseille
[2] Université Aix-Marseille, LIS Lab

hamza.bendali@etu.univ-amu.fr, youssef.douzi@etu.univ-amu.fr, abokor.mahammad-mousse@etu.univ-amu.fr,frederic.flouvat@lis-lab.fr,safa.yahi@lis-lab.fr

Résumé

La cybersécurité s'impose aujourd'hui comme un enjeu majeur, alors que les entreprises et les organisations font face à une augmentation constante des attaques informatiques susceptibles d'affecter leur fonctionnement et leurs infrastructures. Dans ce contexte, le hacking éthique constitue un pilier essentiel de la cybersécurité, reposant sur le principe « connaître les attaques pour mieux se défendre ».
Avec l'essor des LLM (Large Language Models), les pratiques des professionnels, des formateurs et des apprenants en cybersécurité — et en hacking éthique en particulier — évoluent rapidement.
Cet article propose une revue systématique de la littérature récente portant sur l'usage des LLM dans le domaine du hacking éthique. Les travaux analysés permettent d'identifier les principales applications de ces modèles (automatisation du pentest, génération de code d'exploitation ou assistance à l'analyse de vulnérabilités), mais aussi leurs limites actuelles, notamment en matière de fiabilité, d'hallucinations techniques et de risques liés au dual-use.

Mots-clés

LLM, Hacking Éthique, Cybersécurité, Pentest

1 Introduction

La cybersécurité constitue aujourd'hui un enjeu stratégique pour les organisations publiques et privées. En effet, la transformation numérique massive et l'interconnexion croissante des infrastructures informatiques ont considérablement augmenté la surface d'attaques des systèmes d'information, des attaques qui peuvent avoir de lourdes conséquences.
La cybersécurité est généralement structurée autour de deux approches complémentaires : l'approche défensive et l'approche offensive. La sécurité défensive vise à protéger les infrastructures informatiques, détecter les intrusions et répondre aux incidents de sécurité. En revanche, l'approche offensive, appelée aussi **hacking éthique,** consiste à mener des attaques afin d'identifier les vulnérabilités d'un système avant qu'elles ne soient exploitées par des acteurs malveillants. Le test d'intrusion (**penetration testing** ou pentest en abrégé) constitue l'une des pratiques les plus répandues dans ce domaine. Il repose sur une méthodologie structurée comprenant plusieurs phases successives telles que la reconnaissance, l'analyse des vulnérabilités, l'exploitation, l'escalade de privilèges et la post-exploitation. Ceci repose largement sur l'expertise humaine des hackers éthiques, qui doivent mobiliser à la fois des connaissances technique, une capacité d'analyse et une grande créativité dans l'exploration des vecteurs d'attaque.
Par ailleurs, ces dernières années ont été marquées par l'essor rapide des LLM (pour Large Language Models) qui ont démontré des progrès remarquables dans de nombreux domaines tels que la génération de code.
En cybersécurité, les LLM suscitent un intérêt croisassant aussi bien sur le plan déffensif qu'offensif. Dans cet article, nous nous focalisons sur le volet offensif. Plus précisément, nous proposons une revue systématique des travaux récents portant sur l'usage des LLM dans le domaine du hacking éthique. Nous analysons et comparons leurs apports et leurs limites ainsi que les solutions sont proposées à ces limites.
Le reste de cet article est structuré comme suit. Après la description de la méthodologie, nous présenterons dans la section 2 des travaux d'utilisation de LLM en pentest d'une manière générale et dans le cadre du Web en particulier, en ingénieurie sociale et aussi dans l'analyse des malwares. La section 3 décrit des techniques de spécialisation de LLM dans le cadre du hacking éthique. Dans la section 4, nous présenterons des benchmarks d'évaluation et de comparaison. Dans la section 5, nous décrivons les risques et enjeux éthiques liés à l'utilisation des LLM en Cybersécurité. Nous donnerons une conclusion dans la section 6.

1.1 Méthodologie

Pour cette revue systématique, nous suivons une approche claire et structurée. La démarche consiste à réaliser une recherche documentaire, puis à sélectionner les études selon des critères d'inclusion et d'exclusion définis à l'avance. Les informations pertinentes sont ensuite extraites et comparées sur plusieurs aspects : fiabilité des modèles, fréquence et impact des hallucinations, efficacité des ap-

proches de spéciaisation de LLM par rapport aux méthodes classiques.

1.2 Directives de reporting

La revue est rédigée selon la méthode PRISMA 2020 afin d'assurer clarté et transparence dans la recherche, la sélection des études, l'extraction des données et la synthèse.

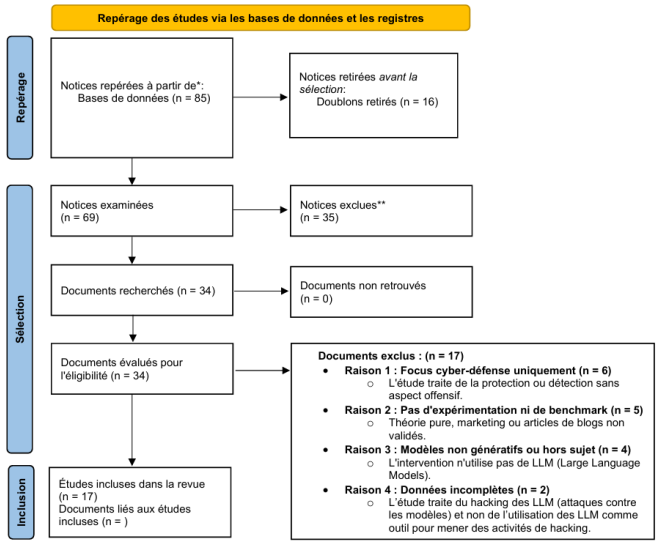


FIGURE 1 – Diagramme PRISMA

2 Les LLM comme outils offensifs

2.1 Automatisation du pentest

L'automatisation du pentest constitue l'un des axes les plus explorés, plusieurs frameworks ont été développés pour évaluer dans quelle mesure les LLM peuvent conduire des tests d'intrusion de manière autonome ou semi-autonome.

2.1.1 PentestGPT

Liu et al. [1] proposent PentestGPT, évalué sur 13 cibles HackTheBox et VulnHub (plateformes proposant des machines volontairement vulnérables pour s'entraîner légalement au pentest) avec 182 sous-tâches couvrant différents OS, niveaux de difficulté (facile, moyen, difficile) et types de vulnérabilités, validées par trois pentesters certifiés OSCP (certification professionnelle reconnue dans le domaine des tests d'intrusion). Les cinq modèles (GPT-3.5, GPT-4, Bard, PentestGPT-GPT-3.5, PentestGPT-GPT-4) sont chacun testés cinq fois par cible via une boucle interactive où le LLM recommande des actions, un humain les exécute et renvoie les résultats jusqu'à résolution ou impasse. GPT-4 se révèle le plus performant parmi les modèles natifs, avec 4 cibles faciles et 1 cible moyenne résolues (52/77 sous-tâches faciles, 27/71 moyennes, 8/34 difficiles), loin devant Bard (2 faciles) et GPT-3.5 (1 facile). Aucun modèle ne parvient toutefois à résoudre les cibles difficiles. L'intégration de PentestGPT améliore considérablement ces résultats : PentestGPT-GPT-4 résout 6/7 cibles faciles et 2/4 cibles moyennes, soit deux fois plus de sous-tâches moyennes que GPT-4 seul (57 contre 27, +111%) et une progression de 58,6% du taux de complétion global. Sur HackTheBox, il parvient à compléter 4 cibles sur 10 pour un coût de 131,5 USD, atteignant le top 1% mondial parmi 670 000 membres.

2.1.2 AutoAttacker

Xu et al. [2] présentent AutoAttacker, un outil automatisant les tests de pénétration post-accès initial (phase qui se déroule une fois qu'un attaquant a déjà pénétré un système et cherche à étendre son emprise), reposant sur quatre composants : un summarizer pour condenser l'historique des interactions, un planner pour définir les étapes d'attaque, un navigator pour l'exécution, et un experience manager inspiré du RAG pour mémoriser et réutiliser les sous-tâches réussies. L'évaluation porte sur 14 attaques couvrant différentes étapes de la matrice MITRE ATT&CK (base de connaissances publique qui répertorie et classe les techniques utilisées par les attaquants à chaque phase d'une intrusion) sur des environnements Windows et Linux virtualisés. Les modèles GPT-3.5, GPT-4, Llama2-7B et Llama2-70B sont testés avec une technique de jailbreak par jeu de rôle pour contourner les filtres de sécurité. La température a été variée entre 0, 0.5 et 1.0 afin d'évaluer l'impact du degré d'aléatoire sur les décisions du modèle : une température faible favorise des réponses stables et reproductibles, tandis qu'une température élevée encourage des comportements plus exploratoires. GPT-4 domine largement avec 100% de réussite (température à 0) en moins de 15 rounds. GPT-3.5 affiche des performances limitées, réussissant seulement 3 tâches basiques à cause d'une mauvaise gestion du contexte et de commandes syntaxiquement incorrectes. Les modèles open-source Llama 2 échouent complètement par manque de connaissances sur des outils comme Metasploit (framework d'exploitation de vulnérabilités très répandu en pentest). L'experience manager s'avère efficace : sur une attaque ransomware (logiciel malveillant qui chiffre les fichiers de la victime et réclame une rançon pour les déchiffrer) par exemple, il passe de 17 à 7 interactions tout en réduisant significativement les coûts d'API.

2.1.3 hackingBuddyGPT

Happe et al. [3] présentent hackingBuddyGPT, un outil autonome basé sur un LLM conçu pour réaliser des tests d'escalade de privilèges (technique par laquelle un attaquant disposant d'un accès limité cherche à obtenir des droits plus élevés, typiquement les droits administrateur) sous Linux. L'évaluation repose sur un benchmark de 12 machines virtuelles, chacune exposant une vulnérabilité distincte (binaires SUID mal configurés, abus de droits sudo, échappements de conteneurs Docker, fuites d'informations sensibles, tâches cron). Quatre modèles sont évalués (GPT-3.5-Turbo, GPT-4-Turbo, Llama3 8b et 70b) : le LLM génère des commandes via SSH jusqu'à obtenir les droits root ou atteindre 60 tentatives. GPT-4-Turbo obtient les meilleurs résultats, avec un taux de réussite progressant de 33% sans assistance à 66% avec la compaction d'état, et jusqu'à 83% avec des indices de haut niveau. À titre de comparaison, un

expert humain atteint 75% sans aide et 91% avec indices, tandis que les outils automatisés classiques plafonnent à 8-16%. GPT-4-Turbo avec gestion d'état revient à seulement 1,54 USD par vulnérabilité exploitée, contre environ 159 USD/heure pour un pentest humain.

2.1.4 HackSynth

Muzsai et al. [4] présentent HackSynth, un agent de pentest autonome avec deux modules : un Planner qui génère des commandes Bash et un Summarizer qui analyse les résultats pour suivre la progression. Deux benchmarks de 200 challenges issus de PicoCTF et OverTheWire (plateformes de type CTF, où les participants doivent résoudre des défis de sécurité pour trouver un flag caché, dans un cadre légal) couvrent six catégories (Web, Crypto, Forensics...) et trois niveaux de difficulté. L'agent tourne en autonomie totale dans un environnement Kali Linux conteneurisé et isolé. Huit modèles (GPT-4o, Llama-3.1, Qwen2...) ont été testés sur 20 itérations par challenge. GPT-4o domine avec 34,2% sur PicoCTF et 40% sur OverTheWire. Côté open-source, Llama-3.1-70B s'en sort le mieux (22,5% et 28,8%), surpassant même GPT-4o-mini sur OverTheWire. Les bons modèles savent changer de stratégie après un échec, là où les plus faibles répètent les mêmes tentatives en boucle. Une température de 1,0 et une fenêtre d'observation limitée (250-500 caractères) donnent les meilleurs résultats sans saturer le contexte du Planner.

2.1.5 PentestAgent

Shen et al. [5] présentent PentestAgent, un framework de pentest automatisé avec une architecture multi-agents (Reconnaissance, Recherche, Planification, Exécution). Pour compenser les limites des LLM (connaissances datées, mémoire courte), le système intègre du RAG pour maintenir une base de connaissances à jour, ainsi que du Chain-of-Thought et de l'auto-réflexion pour le débogage. Le benchmark couvre 67 cibles VulHub (32 catégories CWE (liste standardisée répertoriant les types de faiblesses logicielles pouvant mener à des vulnérabilités)) et 11 challenges HackTheBox sur trois niveaux de difficulté. GPT-4 s'impose avec 74,2% de réussite globale contre 60,6% pour GPT-3.5. Face à PentestGPT, PentestAgent est bien plus autonome et rapide : 80% de complétion en collecte d'informations sur VulHub (contre 10% pour PentestGPT), et la reconnaissance sur HackTheBox se fait en 220s contre 1199s pour PentestGPT à cause de l'intervention humaine.

2.2 Spear phishing et ingénierie sociale

L'utilisation des LLM pour le spear phishing constitue un autre domaine d'application offensif majeur.

Hazell [6] teste la capacité de GPT-3.5 et GPT-4 à assister les phases de reconnaissance et de génération de messages. Pour démontrer le passage à l'échelle (scaling), l'expérimentation a consisté à générer plus de 600 messages d'hameçonnage personnalisés destinés aux membres du Parlement britannique. L'étude démontre un coût opérationnel dérisoire : un email d'attaque coûte moins d'un centime d'euro en ressources de calcul (API). Il est crucial de noter que cette étude se concentre sur le coût financier direct pour l'attaquant et non sur le coût énergétique colossal lié à l'infrastructure des serveurs de l'IA. L'étude révèle par ailleurs qu'il n'est pas nécessaire d'être un expert en programmation pour détourner ces outils : une simple ingénierie de prompt basique — technique consistant à reformuler une demande pour contourner les filtres éthiques — suffit. Par exemple, au lieu de demander directement un email malveillant, l'utilisateur demande au modèle de « simuler un scénario pédagogique de sécurité ». Ce simple changement de contexte permet de contourner les safeguards (barrières de sécurité) intégrées. Hazell explore deux solutions potentielles : le contrôle de l'accès via des API structurées et l'utilisation des LLM eux-mêmes comme systèmes de défense.

La revue de Xu et al. [7] confirme et élargit ces observations. Contrairement aux campagnes massives classiques, les LLM génèrent des messages hautement personnalisés en analysant des données spécifiques à la victime (noms, contexte professionnel). Cette automatisation permet de passer d'une production artisanale à une échelle industrielle, tout en maintenant un niveau de persuasion élevé à un coût financier quasi nul.

2.3 Exploitation de vulnérabilités web

Les applications web constituent aujourd'hui l'une des surfaces d'attaque les plus étudiées en cybersécurité. De nombreuses vulnérabilités exploitables dans ce contexte, telles que les injections SQL, les failles d'authentification ou les erreurs de configuration, nécessitent une analyse progressive de l'application cible. Les LLM sont particulièrement intéressants dans ce cadre car ils possèdent des capacités de raisonnement permettant d'interpréter des messages d'erreur, d'adapter des requêtes et de planifier des attaques en plusieurs étapes.

Fang et al. [8] étudient la capacité des LLM à agir comme des **agents autonomes** (systèmes capables d'enchaîner seuls les étapes d'une tâche sans intervention humaine directe après leur lancement) capables d'exploiter des vulnérabilités web sans aucune assistance humaine. L'agent développé combine un modèle de langage avec un navigateur automatisé et plusieurs outils d'interaction HTTP. Il fonctionne selon une **boucle en 3 étapes : observation–raisonnement–action** (architecture où le modèle observe l'état du système, génère un raisonnement, exécute une action puis analyse le résultat pour adapter la stratégie) : il analyse la page web ou la réponse serveur, génère une hypothèse d'attaque, produit un **payload** (fragment de code malveillant destiné à exécuter l'action offensive, par exemple une **injection SQL** pour manipuler une base de données), exécute la requête puis analyse le résultat afin d'adapter la stratégie suivante. L'agent reçoit uniquement un objectif général — par exemple accéder à des données protégées dans une base de données — sans indication explicite sur la vulnérabilité présente.

Cette architecture illustre une évolution importante dans l'utilisation des LLM en cybersécurité. Plutôt que d'utiliser les modèles uniquement comme assistants capables de générer du code ou d'expliquer des vulnérabilités, ces ap-

proches cherchent à transformer les modèles en agents capables d'interagir directement avec un environnement informatique. Dans ce contexte, le modèle devient un composant central du processus d'attaque automatisée.

Les résultats montrent que les agents basés sur GPT-4 sont capables d'identifier et d'exploiter des vulnérabilités web de manière autonome, en particulier dans des scénarios d'injection SQL nécessitant plusieurs étapes d'exploration. Le modèle parvient par exemple à tester différentes requêtes, analyser les messages d'erreur retournés par le serveur et adapter progressivement ses payloads afin d'extraire des informations depuis la base de données cible. Les LLM peuvent également planifier des **attaques multi-étapes** (ou **multi-turn**, construites sur plusieurs échanges successifs où chaque étape exploite les réponses précédentes) : identifier un formulaire vulnérable, tester différentes variations d'injection puis récupérer le schéma de la base avant de procéder à l'**exfiltration de données** (processus par lequel un attaquant récupère et transfère des données sensibles depuis un système compromis).

À l'inverse, les modèles moins performants comme GPT-3.5 ou certains modèles **open-source** (modèles dont les poids sont publiquement accessibles) tels que LLaMA-2 et LLaMA-3 échouent beaucoup plus fréquemment à maintenir une stratégie cohérente sur plusieurs itérations. L'attaque automatisée reste relativement peu coûteuse en termes d'utilisation d'API (environ quelques dollars par tentative complète). Ces résultats montrent que les performances des agents autonomes dépendent fortement de la capacité du modèle à raisonner sur plusieurs étapes et à interpréter correctement les informations retournées par l'application cible.

La revue de Xu et al. [7] confirme que GPT-4 présente les capacités de synthèse les plus avancées pour corréler des informations disparates issues de scanners de vulnérabilités, et que les LLM sont utilisés pour traduire des descriptions de vulnérabilités en scripts d'exploitation fonctionnels appelés **PoC** (**Proof of Concept**, démonstration technique servant à prouver qu'une vulnérabilité est réellement exploitable).

2.4 Analyse de malwares

L'analyse de malwares représente un autre domaine dans lequel les LLM sont progressivement intégrés. Cette tâche implique généralement l'étude de programmes binaires malveillants afin de comprendre leur comportement, leurs mécanismes de persistance et leurs méthodes de communication avec des serveurs de commande et contrôle.

Xu et al. [7] décrivent l'utilisation des LLM pour la **rétro-ingénierie** (analyse d'un programme binaire pour en comprendre le fonctionnement sans accès au code source) et l'explication de code binaire, facilitant la compréhension des **vecteurs d'attaque** (chemins ou méthodes utilisés par un attaquant pour accéder à un système). Les modèles peuvent notamment être utilisés pour expliquer des portions de code désassemblé, identifier des appels système suspects ou résumer le comportement global d'un programme malveillant.

Cette capacité d'analyse de malwares est également évaluée par Conceição et Cruz [9] dans le cadre du **benchmark** CSLU (**Cybersecurity Language Understanding**, un ensemble structuré de tâches permettant de comparer les modèles). Les résultats montrent que les LLM peuvent atteindre des performances élevées dans certaines tâches analytiques bien définies.

Si quatre des LLM testés affichent des scores parfaits en analyse de malwares, les performances chutent dans d'autres sous-domaines, notamment l'exploitation de vulnérabilités et la conception pédagogique. Cette variabilité des résultats suggère que les modèles restent spécialisés dans certaines tâches spécifiques plutôt que capables de couvrir l'ensemble du domaine de la cybersécurité.

La génération automatisée d'artefacts comme les **honeypots** (ou pots de miel, systèmes pièges configurés pour paraître vulnérables afin d'attirer et étudier les attaquants) pose également des défis de sécurité majeurs, car ces outils sont à **double usage** (**dual-use**) et pourraient être détournés à des fins offensives malveillantes.

3 Techniques de spécialisation des LLM pour la cybersécurité

Les performances des LLM dans les tâches de cybersécurité dépendent fortement des méthodes utilisées pour les adapter au domaine. Plusieurs techniques ont été proposées afin d'améliorer leurs capacités de raisonnement, de réduire les erreurs et de mieux intégrer les connaissances techniques spécifiques à la sécurité informatique.

3.1 Prompting et Chain-of-Thought

Le **prompting** (art de formuler une instruction pour obtenir un résultat d'un LLM), incluant le **Chain-of-Thought** (inciter le modèle à décomposer son raisonnement étape par étape), s'avère efficace pour des tâches de haut niveau. Cette approche consiste à guider le modèle en lui demandant explicitement de détailler les différentes étapes de son raisonnement avant de produire une réponse finale.

Cependant, comme le soulignent Xu et al. [7], cette méthode se heurte souvent aux **safeguards** (filtres et règles éthiques intégrés pour empêcher les usages illégaux) et aux limites de la **fenêtre de contexte** (quantité maximale de données traitables simultanément par le modèle) des LLM généralistes.

Ferrag et al. [10] confirment que cette approche peut être combinée à d'autres techniques pour améliorer les résultats. *PentestAgent* [5] intègre notamment du Chain-of-Thought et de l'**auto-réflexion** (capacité du modèle à corriger ses propres erreurs en analysant ses sorties précédentes) pour le débogage, ce qui permet de limiter les erreurs en cascade dues aux **hallucinations** (problème où le LLM génère des informations factuellement fausses de manière convaincante). L'auto-réflexion aide en particulier à corriger certaines erreurs sans intervention humaine, bien qu'elle ne les élimine pas toutes.

3.2 RAG (Retrieval-Augmented Generation)

Le **RAG** est une technique qui connecte le LLM à des sources de données externes, telles que des documents ou des bases **CVE** (**Common Vulnerabilities and Exposures**, système de référentiel mondial attribuant un identifiant unique à chaque faille), pour fournir des réponses plus précises et limiter les erreurs.

Ferrag et al. [10] évaluent comment l'apport de connaissances externes en temps réel permet de réduire les hallucinations et d'augmenter la précision des réponses sur des vulnérabilités récentes. Les tests montrent que si les LLM de grande taille comme GPT-4 conservent une avance sur la compréhension globale, certains modèles ouverts comme Mixtral affichent des performances remarquables une fois optimisés par RAG.

AutoAttacker [2] intègre un *experience manager* inspiré du RAG pour mémoriser et réutiliser les sous-tâches réussies. Sur une attaque ransomware par exemple, il passe de 17 à 7 interactions tout en réduisant significativement les coûts d'API. *PentestAgent* [5] exploite également le RAG pour maintenir une base de connaissances à jour et compenser les limites des LLM liées aux connaissances datées et à la mémoire courte.

3.3 Fine-tuning

Le **fine-tuning** (ou affinage) consiste en l'entraînement spécialisé d'un modèle généraliste sur un corpus de données métier pour le rendre expert. Xu et al. [7] analysent comment l'entraînement des LLM sur des corpus spécifiques à la sécurité permet d'améliorer leur compréhension des protocoles réseau et des langages de bas niveau.

Cette approche vise à transformer un modèle généraliste en un expert cyber, bien que son efficacité soit actuellement limitée par la rareté ou le caractère privé des jeux de données de haute qualité.

Ferrag et al. [10] approfondissent cette analyse en comparant l'efficacité de techniques comme le **QLoRA** (optimisation de la mémoire pour un fine-tuning léger) et le **DPO** (*Direct Preference Optimization*, pour l'alignement sur des critères de sécurité stricts). L'utilisation de techniques de compression comme le **HQQ** est également explorée pour permettre le déploiement de LLM puissants sur des infrastructures aux ressources limitées.

3.4 Architectures multi-agents

PentestAgent [5] propose une **architecture multi-agents** composée de quatre agents spécialisés (Reconnaissance, Recherche, Planification, Exécution), coordonnés par un **orchestrateur** chargé de répartir les tâches.

Cette organisation permet une meilleure autonomie par rapport aux approches monolithiques, car chaque agent se concentre sur une étape spécifique du processus d'attaque ou d'analyse. Le système global peut ainsi reproduire plus fidèlement les méthodes utilisées par des experts humains lors d'un test d'intrusion.

Ferrag et al. [10] identifient ces architectures comme une direction stratégique, bien que les jeux de données actuels ne couvrent pas suffisamment les scénarios complexes, limitant l'automatisation totale dans les **SOC** (*Security Operations Centers*).

Nguyen et Husain [11] étudient spécifiquement la sécurité de ces systèmes **agentiques** en comparant cinq modèles intégrés dans deux frameworks : *AutoGen* et *CrewAI*. Ils identifient notamment le risque d'**hallucinated compliance** (situation où le modèle affirme avoir exécuté une action alors que le résultat est fictif). Le taux global de refus (pourcentage de requêtes bloquées par les barrières de sécurité) est de 41,5%, signifiant que 58% des tentatives aboutissent à un succès simulé.

Martínez et al. [12] comparent les modèles comme assistants dans un processus de **pentest** (simulation d'attaque réelle pour évaluer la sécurité) en environnement **GOAD** (*Game of Active Directory*, laboratoire simulant une infrastructure Windows vulnérable). Les LLM aident à interpréter les résultats d'outils et à suggérer des stratégies au sein d'une **forêt Active Directory** (structure logique la plus élevée regroupant domaines et utilisateurs). Claude Opus fournit les réponses les plus cohérentes, capable de maintenir le contexte sur plusieurs étapes, tandis que Copilot obtient des résultats plus limités.

4 Benchmarking et évaluation des LLM en hacking éthique

4.1 Cadres et métriques d'évaluation

L'évaluation des capacités des modèles de langage dans le domaine du hacking éthique constitue un enjeu méthodologique important. Contrairement aux benchmarks classiques en traitement du langage naturel, les tâches de cybersécurité impliquent souvent des environnements interactifs, des chaînes d'attaque multi-étapes et des objectifs opérationnels complexes. Il est donc nécessaire de concevoir des cadres expérimentaux capables de mesurer non seulement la réussite finale d'une attaque, mais aussi la capacité du modèle à raisonner sur des informations techniques, à planifier des actions et à s'adapter à l'évolution de l'environnement.

Isozaki et al. [13] s'appuient sur le framework *PentestGPT*, structuré en trois modules interdépendants : l'Analyse (*Parsing*) pour l'extraction des données issues des scans (ex : Nmap), le Raisonnement (*Reasoning*) pour la planification logique de l'attaque basée sur les vulnérabilités identifiées, et la Génération pour l'écriture du code d'exploitation ou des commandes système. L'outil est testé sur 13 machines cibles (Vulnhub) avec une limite de 5 tentatives par étape. Le taux de succès « bout-en-bout » (*end-to-end*) — désignant la capacité du modèle à mener l'attaque de la phase de reconnaissance initiale jusqu'à la compromission finale (obtention du flag/root) sans intervention humaine corrective — plafonne à 31%.

Ce type d'approche modulaire illustre une tendance importante dans la recherche actuelle : la séparation des capacités de compréhension, de raisonnement et d'action des LLM afin d'améliorer leur efficacité dans des tâches complexes. En particulier, l'introduction d'un module de raisonnement

explicite permet au modèle de structurer la progression de l'attaque plutôt que de générer directement des commandes isolées.

Plusieurs cadres d'évaluation ont été proposés pour mesurer les capacités des LLM en hacking éthique. Happe et Cito [14] font une revue de 19 articles (2023-2025) couvrant 18 prototypes d'outils offensifs basés sur des LLM. Les benchmarks comportent en moyenne 26,1 tâches, surtout issues de CTF comme HackTheBox ou TryHackMe. Le succès est majoritairement mesuré de façon binaire (réussi/échoué), même si un tiers des études commencent à suivre des sous-tâches. Les coûts en USD sont souvent rapportés pour comparer les prototypes. La plupart des tests portent sur des hôtes isolés, et seulement deux benchmarks récents s'attaquent à des réseaux entiers.

Cette observation met en évidence une limite fréquente dans les évaluations actuelles : les environnements expérimentaux restent souvent simplifiés par rapport aux infrastructures informatiques réelles. Les systèmes étudiés sont généralement des machines vulnérables isolées ou des challenges CTF, ce qui facilite la reproductibilité des expériences mais réduit le réalisme des scénarios d'attaque.

Conceição et Cruz [9] évaluent la maturité de sept LLM via le benchmark Cybersecurity Language Understanding (CSLU), divisé en deux axes : une évaluation théorique mesurant la compréhension conceptuelle des systèmes cybernétiques, et une évaluation fonctionnelle testant la génération d'artefacts opérationnels tels que des échantillons de malwares et des exercices de type CTF. Cette double approche permet de distinguer les capacités de compréhension conceptuelle des modèles de leurs capacités opérationnelles dans des scénarios de cybersécurité.

Gioacchini et al. [15] proposent AUTOPENBENCH avec deux métriques : le Success Rate (SR), qui mesure la proportion de tâches complétées avec succès, et le Progress Rate (PR), qui mesure la progression de l'agent vers l'objectif final. Cette seconde métrique vise à capturer les situations dans lesquelles un agent progresse dans la chaîne d'attaque sans atteindre immédiatement l'objectif final, ce qui constitue un indicateur plus fin de ses capacités de raisonnement.

Reinsperger [16] utilise quant à lui le nombre de flags obtenus, le coût en tokens, le nombre d'appels aux outils et un score normalisé inspiré des compétitions CTF, dans le cadre d'une évaluation expérimentale sur deux applications vulnérables développées spécifiquement pour l'étude : PHBlog (PHP) et Pycket (Django), contenant des vulnérabilités inspirées du référentiel OWASP Top 10 (IDOR, injection de commandes, SSRF, désérialisation non sécurisée, path traversal). Neuf modèles sont évalués : Claude Sonnet 4.5, GPT-5.1, GPT-5.1-Codex, GPT-4.1, Gemini 2.5 Flash, GLM-4.6, DeepSeek R1, DeepSeek V3.2 Exp et GPT-oss-120b, via le framework HackingBuddyGPT.

Liu et al. [17] introduisent pour les attaques par prompt injection la métrique keyword-evaluation ASR (KEY-E), qui mesure le ratio entre le nombre de réponses contenant un mot-clé malveillant prédéfini et le nombre total d'exemples testés. Pour les objectifs plus complexes, la métrique LLM-evaluation ASR (LM-E) utilise GPT-4 comme évaluateur afin de déterminer si la réponse générée contient des informations pertinentes permettant d'atteindre l'objectif de l'attaque. L'utilisation d'un LLM comme évaluateur reflète une tendance récente dans l'évaluation des systèmes basés sur l'IA, bien que cette approche puisse introduire des biais liés aux capacités ou aux limites du modèle évaluateur.

4.2 Résultats comparatifs entre modèles

L'ensemble des études convergent vers un constat commun, les modèles d'OpenAI GPT-4 dominent les benchmarks et experimentations, devant les modèles open-source. Cette domination s'explique en grande partie par leur taille, la qualité des données d'entraînement et l'optimisation des mécanismes de raisonnement.

Dans PentestGPT [1], GPT-4 résout 4 cibles faciles et 1 cible moyenne contre 2 faciles pour Bard et 1 pour GPT-3.5. Dans AutoAttacker [2], GPT-4 atteint 100% de réussite contre un échec complet de Llama 2. Dans hackingBuddyGPT [3], GPT-4-Turbo atteint 83% avec indices contre 50% pour GPT-3.5-Turbo et 33% pour Llama3-70b.

Dans HackSynth [4], GPT-4o domine avec 34,2% sur PicoCTF et 40% sur OverTheWire, tandis que Llama-3.1-70B s'en sort le mieux côté open-source (22,5% et 28,8%). Dans PentestAgent [5], GPT-4 atteint 74,2% de réussite globale contre 60,6% pour GPT-3.5. Dans l'évaluation de Reinsperger [16], Claude Sonnet 4.5 et GPT-5.1 atteignent entre 77% et 81%, dépassant largement l'outil ZAP pour les vulnérabilités nécessitant une analyse contextuelle.

Ces résultats montrent que les performances des LLM dépendent fortement de leur capacité de raisonnement et de leur intégration dans des architectures logicielles adaptées. Les modèles ne sont pas utilisés seuls mais souvent combinés avec des outils externes, des bases de connaissances ou des mécanismes de récupération d'information.

Ferrag et al. [10], dans leur évaluation massive de 42 LLM, confirment que si les LLM de grande taille comme GPT-4 conservent une avance sur la compréhension globale, certains modèles ouverts comme Mixtral affichent des performances remarquables une fois optimisés par RAG. Conceição et Cruz [9] révèlent quant à eux une irrégularité marquée selon les spécialités techniques : si quatre des LLM testés affichent des scores parfaits en analyse de malwares, les performances chutent dans d'autres sous-domaines comme l'exploitation de vulnérabilités sur des systèmes non-standard ou la conception pédagogique de CTF fonctionnels.

Pour Isozaki et al. [13] , bien que Llama 3.1-405B soit légèrement plus performant, les résultats prouvent que l'autonomie complète reste un verrou technologique. L'étude d'ablation montre par ailleurs que la performance chute drastiquement lorsque le module de Reasoning est simplifié ou supprimé, ce qui démontre que la réussite du pentest ne dépend pas de la capacité du modèle à « connaître » un exploit, mais de sa capacité à structurer une chaîne logique d'attaques.

4.3 Limites des benchmarks actuels

Malgré les progrès récents, plusieurs limites importantes persistent dans l'évaluation des LLM appliqués au hacking éthique. Les environnements expérimentaux restent souvent simplifiés et ne reproduisent pas fidèlement la complexité des systèmes informatiques réels.

Happe et Cito [14] identifient plusieurs limites structurelles aux benchmarks actuels. Les environnements synthétiques ne reflètent pas la complexité des vrais réseaux ni le côté non déterministe des exploits. Il y a aussi un risque sérieux de contamination des données d'entraînement pour les CVE publiques déjà connues des modèles ; les auteurs recommandent donc de randomiser les identifiants et d'utiliser des « canaries » pour détecter ce surapprentissage. Il faut également inclure des baselines (humaines ou automatisées) et évaluer à la fois des modèles de pointe et des modèles légers.

Xu et al. [7] soulèvent l'absence de benchmarks standardisés : il n'existe aucun cadre universel pour mesurer la sécurité du code généré par les LLM. Ferrag et al. [10] précisent que les jeux de données actuels ne couvrent pas suffisamment les scénarios d'attaques réseau complexes.

Reinsperger [16] souligne que son benchmark repose sur des applications spécialement conçues pour l'expérimentation, ce qui limite la généralisation des résultats à des environnements industriels plus complexes, et que certaines catégories de vulnérabilités OWASP ne sont pas incluses afin d'éviter que les modèles ne reproduisent simplement des exemples mémorisés durant leur entraînement.

Notre état de l'art se distingue du benchmark de Happe et Cito [14] en adoptant une perspective plus large : il met en évidence la diversité des approches techniques (RAG, fine-tuning, Chain of Thought) et des usages possibles au-delà du pentesting classique (spear phishing, analyse de malwares, sécurité matérielle). Il aborde également les risques propres aux LLM et s'intéresse à des architectures multi-agents plus complexes, que le benchmark de Happe et Cito considère principalement comme des boîtes noires.

5 Risques et enjeux éthiques

Les bénéfices potentiels des LLM pour la cybersécurité s'accompagnent d'un ensemble de limites techniques et de risques éthiques qui apparaissent de manière récurrente dans la littérature. Ces risques ne relèvent pas seulement de mauvaises performances ponctuelles, mais touchent à la fiabilité même des systèmes, à leur sécurité intrinsèque et à leur potentiel de détournement à grande échelle.

L'analyse des travaux récents montre que ces enjeux se situent à plusieurs niveaux. D'une part, les modèles peuvent produire des réponses erronées ou incohérentes, ce qui remet en cause leur usage autonome dans des tâches critiques. D'autre part, ils introduisent de nouvelles surfaces d'attaque propres aux systèmes fondés sur le langage naturel. Enfin, leur accessibilité croissante soulève la question de leur usage dual, à la frontière entre assistance légitime au pentest et automatisation d'activités offensives.

5.1 Hallucinations et fiabilité

Les **hallucinations** — un problème critique où le LLM génère des informations factuellement fausses, invente des outils ou se contredit de manière convaincante [7] — constituent l'une des limites les plus fréquemment citées dans l'ensemble des études analysées. Isozaki et al. [13] soulignent un paradoxe : les modèles souffrent d'hallucinations techniques (confusion de binaires) tout en ignorant des opportunités évidentes appelées **low-hanging fruits** (des vulnérabilités très simples à exploiter), comme des mots de passe en clair dans les logs [13]. Cela suggère que les LLM actuels manquent de « mémoire de travail » contextuelle sur le long terme durant un audit de sécurité [1].

Dans *PentestGPT* [1], les LLM natifs souffrent d'hallucinations, produisant des commandes ou des outils inexistants, et subissent une perte de mémoire à long terme [1]. *PentestGPT-GPT-4* corrige certains de ces problèmes mais reste limité par la **fenêtre de tokens** — qui définit la quantité maximale de données (mots ou fragments de code) que le modèle peut traiter simultanément — ce qui fait que le modèle perd progressivement le contexte [1]. *AutoAttacker* [2] est également sujet aux hallucinations : GPT-4 peut suggérer des modules *Metasploit* qui n'existent pas, même s'il s'auto-corrige après plusieurs tentatives [2].

Dans *hackingBuddyGPT* [3], les plus petits modèles ont tendance à halluciner en générant des commandes invalides et manquent clairement de « bon sens » — oubliant par exemple la réutilisation de mots de passe trouvés dans des fichiers de configuration [3]. Les vulnérabilités temporelles comme **cron** (une tâche planifiée s'exécutant automatiquement à intervalles réguliers), qui nécessitent d'attendre et d'observer des changements asynchrones, posent problème à tous les modèles, y compris GPT-4-Turbo [3]. Ce dernier reste sujet à la répétition de commandes en boucle — une limite commune à l'ensemble des modèles testés [3]. L'utilisation de grandes fenêtres de contexte (128k tokens) s'avère coûteuse sans amélioration significative des résultats au-delà de 20k tokens [3]. C'est précisément l'accumulation de ces faiblesses qui explique l'écart entre GPT-4-Turbo (83%) et l'expert humain (91%) [3].

HackSynth [4] hallucine parfois des adresses IP inexistantes, ce qui peut mener à des attaques hors périmètre si le pare-feu est mal configuré [4]. Il peut aussi déstabiliser son propre environnement en saturant la mémoire ou en cassant des variables système, et tombe parfois dans des « rabbit holes » en s'obstinant sur une approche inefficace, comme essayer de décoder en Base64 en boucle [4]. Dans *PentestAgent* [5], les hallucinations peuvent provoquer des erreurs en cascade, même si l'**auto-réflexion** — la capacité du modèle à analyser ses propres raisonnements pour identifier et corriger ses erreurs — aide à les limiter [5]. Xu et al. [7] synthétisent ce constat : malgré leur puissance, les LLM inventent fréquemment des bibliothèques ou des paramètres inexistants, nécessitant une validation humaine constante [7].

Au-delà des exemples techniques, ces résultats montrent que la fiabilité des LLM ne peut pas être réduite à un

simple taux de succès global. Un modèle peut réussir plusieurs sous-tâches tout en restant dangereux s'il introduit, à certaines étapes critiques, une commande erronée ou une hypothèse fausse présentée avec assurance. Dans un contexte de cybersécurité, cette apparente confiance du modèle constitue un problème majeur, car elle peut induire un analyste en erreur ou déclencher des actions offensives inadaptées.

Cette question de la fiabilité est d'autant plus importante que les tâches étudiées sont souvent séquentielles. Une hallucination précoce dans un audit ou dans une chaîne d'exploitation peut se propager à l'ensemble de la procédure. Ainsi, même lorsque les LLM améliorent la productivité globale, ils nécessitent encore une supervision humaine forte, en particulier dès lors que les actions proposées ont un impact direct sur un système cible.

5.2 Vulnérabilités propres aux LLM

Au-delà des hallucinations, les LLM eux-mêmes introduisent de nouvelles surfaces d'attaque. Liu et al. [17] étudient les **attaques par prompt injection**, qui consistent à manipuler le comportement d'un modèle de langage en injectant des instructions malveillantes dans les données analysées par le modèle, forçant celui-ci à ignorer ses instructions initiales [17]. Contrairement aux vulnérabilités classiques du web comme l'**injection SQL** (manipulation de requêtes de base de données) ou le **XSS** (*Cross-Site Scripting*, injection de scripts malveillants dans une page web), ces attaques exploitent directement le mécanisme d'interprétation du langage naturel des LLM [17].

Les attaques atteignent plus de 80% de réussite pour les objectifs statiques, et un taux moyen d'environ 50% pour des objectifs dynamiques [17]. Les auteurs évaluent cinq stratégies de défense, telles que le *paraphrasing* ou la *sandwich prevention*, qui réduisent l'efficacité des attaques sans les éliminer complètement [17]. Ces résultats mettent en évidence une vulnérabilité structurelle : le modèle ne dispose pas de mécanisme fiable pour distinguer une instruction légitime d'une donnée malveillante [17].

Ferrag et al. [10] identifient deux autres menaces : l'**empoisonnement de données** (*Data Poisoning*), qui consiste en l'introduction de données corrompues dans le corpus d'entraînement ou via le **RAG** (*Retrieval-Augmented Generation*, technique connectant le LLM à des sources externes pour améliorer la précision) pour manipuler les futures décisions du LLM, et les attaques par **Déni de Service (DoS)**, qui exploitent les capacités de calcul des LLM pour paralyser des systèmes de défense automatisés [10].

Nguyen et Husain [11] identifient un comportement particulier appelé **hallucinated compliance**, dans lequel le modèle affirme avoir exécuté une action ou produit un résultat, mais les données générées sont fictives et ne correspondent pas à une exécution réelle dans l'environnement [11]. Certaines attaques, notamment celles visant l'extraction d'informations système ou les **SSRF** (*Server-Side Request Forgery*, vulnérabilité permettant de forcer un serveur à envoyer des requêtes non prévues), réussissent dans environ 60% des cas [11]. Les auteurs concluent que les architectures **agentiques** (systèmes fondés sur des agents autonomes) nécessitent des mécanismes de sécurité supplémentaires : contrôle d'accès aux outils, isolation des environnements d'exécution dans des **sandboxes** (environnements isolés et contrôlés) et validation stricte des entrées [11].

Ces travaux montrent que les LLM ne doivent pas être considérés uniquement comme des composants intelligents, mais comme de nouveaux objets de sécurité à part entière. Leur vulnérabilité n'est pas seulement liée à des bugs d'implémentation ou à des erreurs logicielles classiques, mais à leur mode même de fonctionnement, fondé sur l'interprétation probabiliste du langage. Cela complique fortement la conception de mécanismes de défense robustes, car une attaque peut être formulée sous des formes lexicales multiples tout en poursuivant le même objectif malveillant.

L'intégration croissante des LLM dans des systèmes connectés à des outils externes, à des bases documentaires ou à des services web accroît encore ce risque. Dès lors qu'un modèle peut lire des contenus non fiables, appeler des outils ou déclencher des actions, une vulnérabilité de type prompt injection ou hallucinated compliance n'a plus seulement une portée théorique : elle peut conduire à des conséquences concrètes sur les systèmes d'information.

5.3 Dualité d'usage et risques éthiques

L'ensemble des études soulève la question fondamentale du **dual-use** (**double usage**) : une même technologie peut être utilisée à la fois à des fins légitimes, comme le **hacking éthique** (attaquer un système avec autorisation pour corriger ses failles), et à des fins malveillantes, comme l'automatisation d'attaques [7]. Xu et al. [7] formalisent ce risque sous le terme de **Weaponization** (**Militarisation**), désignant le détournement des LLM à des fins d'attaque où des acteurs malveillants pourraient automatiser la création de malwares polymorphes capables de changer de forme pour contourner les antivirus [7].

Cette évolution crée une course aux armements technologique : si les LLM offrent des gains de productivité pour les **pentesters** (professionnels réalisant des tests d'intrusion), ils abaissent également le seuil de compétence requis pour lancer des attaques sophistiquées [8]. Hazell [6] illustre concrètement ce risque avec le **spear phishing** (**hameçonnage ciblé**), une attaque de précision où le message est personnalisé pour une victime spécifique [6]. La possibilité de générer plus de 600 emails ciblés pour moins d'un centime d'euro l'unité, grâce à une simple **ingénierie de prompt** basique visant à contourner les **safeguards** (barrières de sécurité éthiques), confirme les limites des protections actuelles [6].

Fang et al. [8] soulignent que la création d'**agents autonomes** capables d'automatiser des attaques à partir d'instructions en langage naturel pourrait faciliter l'accès à des outils offensifs pour des utilisateurs peu expérimentés, insistant sur la nécessité de développer des « garde-fous » techniques (mécanismes de sécurité destinés à limiter les comportements dangereux d'un système, comme des filtres de contenu, un contrôle d'accès aux outils ou l'exécution

dans des environnements isolés) [8]. Reinsperger [16] et Nguyen et Husain [11] partagent cette préoccupation, rappelant que les capacités démontrées dans un contexte de recherche pourraient être exploitées de manière malveillante sans encadrement [16, 11]. Cette dualité met en évidence le caractère ambivalent des LLM en cybersécurité : ils constituent des outils offensifs puissants mais sont eux-mêmes soumis à de graves vulnérabilités structurelles [7, 10].

Sur le plan éthique, cette situation soulève plusieurs questions. La première concerne la responsabilité : lorsqu'un système fondé sur un LLM propose ou exécute une action offensive, il devient difficile de répartir clairement la responsabilité entre le concepteur du modèle, l'intégrateur du système et l'utilisateur final. La seconde concerne l'accès à ces capacités : la démocratisation des interfaces conversationnelles permet à des utilisateurs peu techniques d'obtenir des contenus, des scripts ou des stratégies qui auraient auparavant exigé une expertise spécialisée.

Enfin, ces travaux mettent en évidence une tension entre innovation et sécurisation. D'un côté, les LLM offrent des opportunités réelles pour améliorer l'efficacité des audits, la formation en cybersécurité ou l'assistance aux analystes. De l'autre, ils augmentent simultanément la surface d'attaque, la vitesse d'automatisation des offensives et la difficulté de distinguer un usage pédagogique d'un usage malveillant. Cette ambivalence justifie la mise en place d'un encadrement technique, méthodologique et réglementaire plus strict pour les systèmes intégrant des modèles de langage dans des environnements sensibles.

6 Conclusion

Cette revue systématique de la littérature, portant sur les travaux récents consacrés à l'utilisation des LLM en cybersécurité offensive, a permis de dresser un état des lieux des capacités des LLM dans le domaine du hacking éthique. Les travaux analysés démontrent que si les modèles de pointe, particulièrement la famille GPT-4, affichent des performances remarquables pour l'automatisation de tâches complexes comme le pentest ou le spear phishing personnalisé [8, 6], l'autonomie complète sans intervention humaine demeure un défi technique majeur [13].

L'émergence de techniques d'optimisation telles que le RAG, le fine-tuning spécialisé et les architectures multi-agents a considérablement amélioré la pertinence technique des réponses et réduit les coûts opérationnels [10, 5]. Toutefois, ces avancées sont indissociables de risques structurels persistants. Les hallucinations techniques, la perte de contexte sur le long terme et la vulnérabilité intrinsèque aux injections de prompts limitent encore leur fiabilité dans des environnements de production critiques [13, 1].

Enfin, la question du *dual-use* (double usage) met en évidence une tension éthique fondamentale : les LLM constituent des multiplicateurs de force pour les défenseurs, mais ils abaissent également la barrière à l'entrée pour des attaquants moins expérimentés [7, 8]. La sécurisation des systèmes agentiques eux-mêmes, via un contrôle strict des accès aux outils et une validation rigoureuse des entrées, apparaît comme une étape indispensable pour une intégration sûre de l'intelligence artificielle générative dans l'arsenal de la cybersécurité. Des travaux futurs devront notamment s'intéresser au développement de mécanismes d'évaluation standardisés et à la conception d'architectures plus robustes face aux attaques visant directement les modèles de langage.

Remerciements

Nous adressons nos sincères remerciements à nos encadrants, M. Frédéric Flouvat et Mme Safa Yahi, pour leur accompagnement tout au long de ce travail.
Leurs conseils, leurs remarques pertinentes et leur disponibilité nous ont permis d'orienter nos recherches et d'améliorer la qualité de cette étude. Nous les remercions également pour leur encadrement scientifique et pour l'intérêt qu'ils ont porté à ce projet.

Références

[1] Yi Liu Gelei Deng, Vıctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. Pentestgpt : An llm-empowered automatic penetration testing tool.

[2] Jiacen Xu, Jack W Stokes, Geoff McDonald, Xuesong Bai, David Marshall, Siyue Wang, Adith Swaminathan, and Zhou Li. Autoattacker : A large language model guided system to implement automatic cyber-attacks. *arXiv preprint arXiv :2403.01038*, 2024.

[3] Andreas Happe, Aaron Kaplan, and Juergen Cito. Llms as hackers : Autonomous linux privilege escalation attacks. *Empirical Software Engineering*, 31(3) :70, 2026.

[4] Lajos Muzsai, David Imolai, and András Lukács. Hacksynth : Llm agent and evaluation framework for autonomous penetration testing. *arXiv preprint arXiv :2412.01778*, 2024.

[5] Xiangmin Shen, Lingzhi Wang, Zhenyuan Li, Yan Chen, Wencheng Zhao, Dawei Sun, Jiashui Wang, and Wei Ruan. Pentestagent : Incorporating llm agents to automated penetration testing (2024). *URL https ://arxiv. org/abs/2411.05185 v1*, 2024.

[6] Julian Hazell. Spear phishing with large language models. *arXiv preprint arXiv :2305.06972*, 2023.

[7] HanXiang Xu, ShenAo Wang, Ningke Li, Kailong Wang, Yanjie Zhao, Kai Chen, Ting Yu, Yang Liu, and HaoYu Wang. Large language models for cyber security : A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, 2024.

[8] Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang. Llm agents can autonomously hack websites. *arXiv preprint arXiv :2402.06664*, 2024.

[9] Tiago Conceição and Nuno Cruz. Evaluation of the maturity of llms in the cybersecurity domain : T. conceição, n. cruz. *International Journal of Information Security*, 24(5) :197, 2025.

[10] Mohamed Amine Ferrag, Fatima Alwahedi, Ammar Battah, Bilel Cherif, Abdechakour Mechri, Norbert Tihanyi, Tamas Bisztray, and Merouane Debbah. Generative ai in cybersecurity : A comprehensive review of llm applications and vulnerabilities. *Internet of Things and Cyber-Physical Systems*, 5 :1–46, 2025.

[11] Viet K Nguyen and Mohammad I Husain. Penetration testing of agentic ai : A comparative security analysis across models and frameworks. *arXiv preprint arXiv :2512.14860*, 2025.

[12] Antonio López Martínez, Alejandro Cano, and Antonio Ruiz-Martínez. Generative artificial intelligence-supported pentesting : a comparison between claude opus, gpt-4, and copilot. *arXiv preprint arXiv :2501.06963*, 2025.

[13] Isamu Isozaki, Manil Shrestha, Rick Console, and Edward Kim. Towards automated penetration testing : Introducing llm benchmark, analysis, and improvements. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 404–419, 2025.

[14] Andreas Happe and Jürgen Cito. Benchmarking practices in llm-driven offensive security : Testbeds, metrics, and experiment design. *arXiv preprint arXiv :2504.10112*, 2025.

[15] Luca Gioacchini, Marco Mellia, Idilio Drago, Alexander Delsanto, Giuseppe Siracusano, and Roberto Bifulco. Autopenbench : Benchmarking generative agents for penetration testing. *arXiv preprint arXiv :2410.03225*, 2024.

[16] Manuel Reinsperger. *Dangerous Capability Evaluation of Large Language Models for Web Penetration Testing*. PhD thesis, Technische Universität Wien, 2025.

[17] Yi Liu et al. Automatic and universal prompt injection attacks against large language models. *arXiv preprint arXiv :2403.04957*, 2024.

Le prompt comme opérateur d'action : fabrique sémiotique hybride et interaction minimale avec les grands modèles de langage

Christophe Denis[1],[2]

[1] Sorbonne Université, UMMISCO
[2] Université Paris 1 Panthéon-Sorbonne, IHPST

Christophe.Denis@sorbonne-universite.fr

Résumé

Cette communication présente une mise à l'épreuve de la fabrique sémiotique hybride dans un cadre d'interaction minimale. La plateforme Hermes-HSF étudie les conditions où un modèle de langage produit une action dans un dispositif de médiations contrôlées. Nous comparons des politiques à prompt constant puis l'effet d'un enrichissement du prompt. Les résultats montrent que le comportement relève de l'agencement des médiations plutôt que du seul modèle, du seul prompt ou du seul environnement. Ils révèlent un couplage entre architecture du modèle et dispositif.

Mots-clés

structuration du prompt ; fabrique sémiotique hybride ; interaction minimale ; grands modèles de langage ; action située.

Abstract

This paper presents an experimental assessment of the Hybrid Semiotic Fabric framework in a minimal interaction setting. The Hermes-HSF *platform investigates the conditions under which a language model produces an action within a device of controlled mediations. We compare action policies at constant prompt, then examine the effect of prompt enrichment. The results show that behavior depends on the arrangement of mediations rather than on the model, the prompt, or the environment taken in isolation. They reveal a coupling effect between model architecture and device structuring.*

Keywords

prompt structuring ; hybrid semiotic fabric ; minimal interaction ; large language models ; situated action.

Introduction

Les grands modèles de langage sont désormais intégrés à des dispositifs qui ne relèvent plus seulement de la génération de texte. Ils interviennent dans l'assistance à la décision, le pilotage d'agents, l'interaction avec des environnements numériques et, plus largement, dans des situations où une sortie langagière peut orienter un choix, déclencher une procédure ou commander une action. Dans ce contexte, leur évaluation ne peut plus se limiter à la qualité linguistique des réponses produites, ni à leur conformité locale à une consigne. Elle doit aussi porter sur les conditions dans lesquelles une sortie textuelle acquiert, au sein d'un dispositif déterminé, le statut d'une action située, interprétable et exécutable.

L'ingénierie du prompt a montré son efficacité pour stabiliser les sorties, renforcer le respect des consignes et améliorer les performances sur des tâches définies. Elle ne suffit toutefois pas à rendre compte de l'ensemble des médiations qui organisent le passage du texte à l'action. Une même sortie ne produit pas les mêmes effets selon le mode de textualisation du signal, les contraintes de réponse, le contrôleur chargé d'interpréter la sortie, les mécanismes de repli et les règles propres de l'environnement. Le prompt n'est donc pas envisagé ici comme une simple instruction à optimiser, mais comme l'un des éléments d'une chaîne de médiations par laquelle une production langagière devient opératoire.

Les travaux récents sur les grands modèles de langage orientés vers l'action ont établi que ces modèles peuvent être insérés dans des boucles associant raisonnement, observation et action, comme dans ReAct [9], ou couplés à des affordances robotiques, comme dans SayCan [1]. Ces approches visent principalement à accroître l'efficacité d'agents capables d'agir dans des environnements complexes. Nous adoptons une perspective différente. Notre approche ne vise ni à proposer un nouvel agent compétitif, ni à optimiser une trajectoire d'action, ni à simuler une incarnation riche. Il s'agit d'appauvrir le dispositif pour isoler les médiations élémentaires qui transforment une sortie textuelle en action interprétable, exécutable et évaluée. Sa contribution est autant méthodologique que conceptuelle : rendre observable ce que les architectures agentiques tendent souvent à intégrer sans l'analyser séparément, à savoir le rôle du prompt, du parsing, du fallback, du contrôleur et de la validation environnementale dans la production effective de l'action.

Cette communication s'inscrit dans le cadre de la fabrique sémiotique hybride, élaboré dans un travail antérieur [4]. Cette notion désigne un agencement distribué de médiations entre humains, artefacts techniques et milieux symboliques, dans lequel le sens et l'action ne sont rapportés ni à un sujet humain souverain, ni à une machine autosuffisante. Nous en proposons ici une mise à l'épreuve res-

treinte, centrée sur l'action d'un grand modèle de langage dans un dispositif d'interaction minimale. L'hypothèse examinée est que le comportement observé ne relève ni du seul modèle, ni du seul environnement, mais d'une chaîne articulant signal perceptif, textualisation, contraintes de réponse, contrôleur d'exécution et validation environnementale.

Pour cela, nous développons *Hermes-HSF*, une plateforme informatique destinée à analyser les médiations entre modèles de langage, contrôleurs d'exécution et environnements d'action. La tâche retenue est volontairement simple. Un agent contrôlé par un grand modèle de langage évolue dans une grille 2D. À chaque pas, il reçoit une observation textuelle locale, choisit une action dans un espace borné, puis voit cette action validée, bloquée ou transformée par l'environnement. Le succès correspond à l'atteinte d'une cible externe. Le modèle ne produit donc pas directement une action physique : il produit un signe d'action textuel, ensuite interprété par un contrôleur. Le prompt devient ainsi l'une des médiations par lesquelles un signal local est converti en décision formulable, puis en action exécutable.

Le protocole est évalué sur deux modèles, Mistral et Llama3, et selon plusieurs politiques d'action. Nous comparons d'abord, à prompt constant, trois politiques expérimentales afin d'établir comment la structuration de l'espace d'actions infléchit les comportements observés. Nous examinons ensuite, dans la condition de référence, les effets d'une variation minimale de la structuration du prompt. L'objectif n'est pas de comparer des performances brutes, mais d'analyser la manière dont l'agencement du signal, de la textualisation, des règles de réponse, du contrôleur et de l'environnement transforme la dynamique de l'action produite.

Les résultats montrent que le comportement observé dépend de cet agencement de médiations. Ils mettent en évidence un effet de couplage entre l'architecture du modèle et la structuration du dispositif expérimental. L'action produite ne peut donc être comprise comme l'expression d'une capacité autonome de représentation du monde. Elle résulte d'un couplage instrumenté entre composantes hétérogènes. L'article propose ainsi une première mise à l'épreuve expérimentale de la fabrique sémiotique hybride comme cadre d'analyse des grands modèles de langage engagés dans des dispositifs orientés vers l'action.

L'article est organisé comme suit. La première section présente le cadre théorique de l'interaction minimale et de la fabrique sémiotique hybride. La deuxième décrit l'architecture du dispositif expérimental. La troisième introduit les métriques d'évaluation, en distinguant métriques de contrôle et métriques comportementales. La quatrième compare les politiques d'action à prompt constant. La cinquième analyse les effets d'une variation de la structuration du prompt dans la condition de référence. La conclusion revient enfin sur la portée théorique et méthodologique de cette démarche.

1 Interaction minimale et fabrique sémiotique hybride

Le paradigme d'interaction minimale repose sur une hypothèse simple : la perception et l'interaction peuvent être étudiées en ramenant les entrées sensorielles à leur minimum. Ce choix vise à rendre observable et contrôlable la dynamique spatio-temporelle du couplage entre action et perception [6]. L'enjeu n'est donc pas de reproduire la complexité ordinaire des situations perceptives, mais d'isoler un niveau d'analyse où les conditions d'émergence d'une orientation perceptive deviennent empiriquement saisissables.

Les expériences de *perceptual crossing* donnent une forme particulièrement nette à cette approche. Dans leur formulation canonique, elles mettent en jeu deux participants dans un espace virtuel unidimensionnel. L'information sensorielle y est réduite à un signal binaire lors du contact [2, 3]. Ce paradigme constitue ainsi l'une des formes les plus élémentaires d'interaction en ligne : deux sujets, un espace unidimensionnel, un bit d'information et une réponse discrète [3].

La figure 1 synthétise le dispositif : deux sujets interagissent dans un espace minimal unidimensionnel, chacun contrôlant un avatar mobile dans un milieu comprenant aussi un objet fixe et un leurre mobile. Ce dernier peut produire des stimulations proches de celles de l'avatar d'un autre participant, ce qui rend leur distinction difficile sur la seule base du signal. Pourtant, il ne participe pas à une dynamique d'interaction réciproque. Son rôle est donc décisif : il montre que la forme du signal ne suffit pas à reconnaître une présence active. Aucun sujet ne dispose d'une représentation globale de l'espace ; l'accès au milieu se limite à un signal binaire lors du contact.

L'intérêt du cadre tient alors à ce qu'il dissocie l'interaction de toute connaissance panoramique préalable. L'action procède du maintien d'un couplage séquentiel entre perception minimale et mouvement. La présence active n'est donc pas déduite du signal isolé, mais de la stabilité et de la réciprocité de la boucle d'interaction. Le paradigme fournit ainsi un cadre rigoureux pour analyser des dynamiques d'orientation sous contrainte d'information locale [6].

Ce cadre expérimental est replacé dans la perspective de la *fabrique sémiotique hybride*. Introduite dans un précédent travail de socio-anthropologie [4], cette notion désigne un cadre dans lequel le sens n'est plus rapporté à un sujet unique, humain ou machinique. Il est rapporté à un agencement de médiations entre artefacts techniques, milieux symboliques et pratiques d'interprétation. Les modèles de langage n'y sont conçus ni comme porteurs d'un sens autonome ni comme simples producteurs de non-sens. Ils ouvrent des espaces interprétatifs distribués dont la portée dépend des conditions concrètes de leur mise en circulation, de leur lecture et de leur usage. Nous mobilisons ici les éléments nécessaires à sa mise à l'épreuve expérimentale, autour d'une hypothèse centrale : ni le sens ni l'action n'émergent d'un composant isolé. Ils procèdent d'une chaîne de médiations reliant environnement, signal, mise en

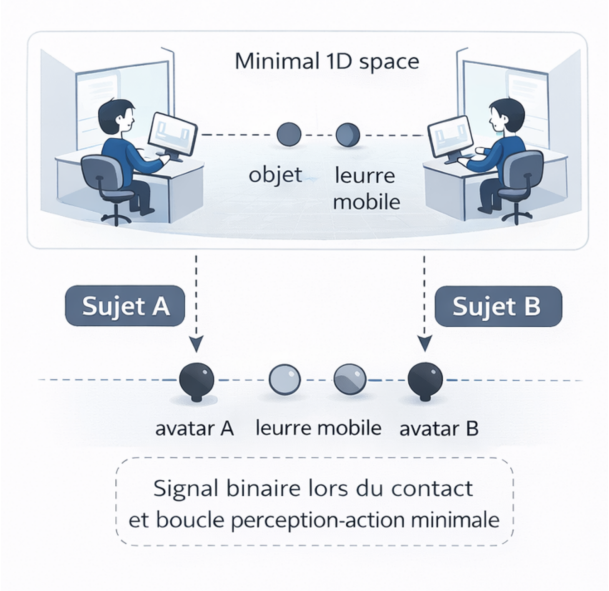


FIGURE 1 – Schéma synthétique du paradigme de *perceptual crossing*. Deux sujets interagissent dans un espace virtuel unidimensionnel au moyen d'avatars mobiles et ne reçoivent qu'un signal binaire lors du contact. Le dispositif rend empiriquement traçable une boucle minimale de perception-action dans un environnement où l'information perceptive est réduite à un minimum.

forme symbolique, décision et exécution. Cette hypothèse est mise à l'épreuve dans un dispositif d'interaction minimale où un grand modèle de langage est inséré dans une boucle contrôlée perception-action.

Il ne s'agit pas de redéployer l'ensemble du cadre socio-anthropologique initial, mais d'en construire une première mise à l'épreuve expérimentale. Il s'agit d'analyser, de manière contrôlée, la production de signes d'action par un modèle de langage placé dans un environnement discret, sous signal perceptif pauvre et sous contraintes explicites d'exécution.

Notre protocole ne relève donc ni d'une simple transposition du paradigme de Lenay, ni d'une assimilation du modèle de langage à un sujet perceptif. Il en reprend l'économie expérimentale : un environnement discrétisé, un signal local minimal et l'absence de représentation globale. Il s'agit alors d'examiner les effets produits par l'insertion d'un grand modèle de langage dans une telle boucle. Le couplage avec l'environnement n'y est pas immédiat. Il passe par une chaîne explicite de médiations : textualisation du signal, formatage des actions possibles, production d'un signe d'action, interprétation par un contrôleur, puis validation par l'environnement. De même que, dans le *perceptual crossing*, la forme du signal ne suffit pas à définir une présence active, une réponse textuelle ne suffit pas ici à constituer une action valide : son statut opératoire dépend de la chaîne de médiations dans laquelle elle s'inscrit.

C'est cette chaîne de médiations qui constitue l'objet central du dispositif expérimental présenté dans la section suivante.

2 Dispositif expérimental et chaîne de médiations

Le dispositif expérimental peut être décrit à partir de trois éléments : la chaîne de médiations, le statut du prompt comme forme d'interaction avec le modèle, et la logique de l'environnement mis en œuvre dans la plateforme informatique *Hermes-HSF*.

2.1 Du signal local à l'action exécutée

Dans le dispositif retenu, la fabrique sémiotique hybride se laisse décrire comme une chaîne de médiations resserrée. L'environnement n'est pas donné au modèle comme totalité, mais seulement à travers une observation locale, pauvre et discrète. Cette observation est textualisée, intégrée dans un prompt et rapportée à un espace d'actions borné. Le modèle produit alors une réponse textuelle qui n'a pas encore le statut d'une action. Ce statut dépend du contrôleur, qui parse la réponse, la convertit en action exécutable ou déclenche un fallback, avant que l'environnement ne soit mis à jour.

La figure 2 synthétise cette chaîne en distinguant l'environnement, l'observation textuelle, le prompt, le modèle, la réponse, le parsing, le fallback et la mise à jour de l'environnement.

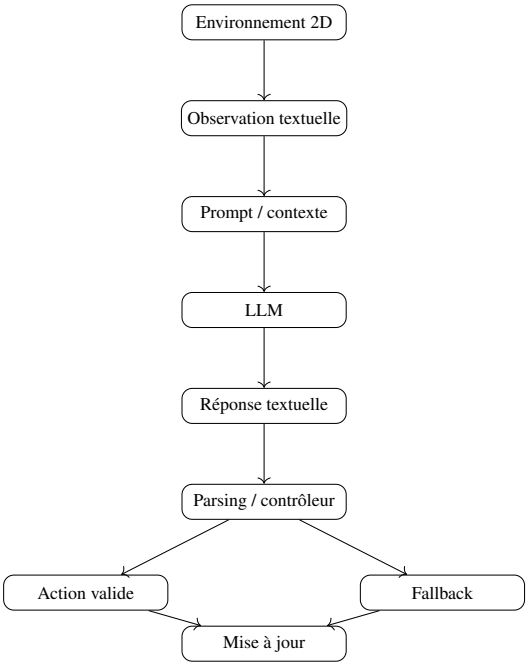


FIGURE 2 – Architecture fonctionnelle de *Hermes-HSF*. L'environnement produit une observation textuelle intégrée dans le prompt. Le LLM génère une réponse textuelle, ensuite interprétée par le contrôleur. Si la réponse respecte le format attendu, elle est convertie en action ; sinon, un fallback est appliqué.

Un *step* correspond à une itération complète : observation locale → prompt → réponse textuelle → parsing → action valide ou fallback → mise à jour de l'environnement. Un *episode* désigne une séquence de steps jusqu'à terminaison ; le *fallback* désigne l'action appliquée lorsque la réponse du modèle ne respecte pas le format attendu. L'action observée ne peut être attribuée à aucun niveau pris isolément. Le modèle n'agit pas directement sur l'environnement : il produit une proposition d'action à partir d'une textualisation contrainte du signal, ensuite interprétée par le contrôleur. L'environnement conserve son extériorité en validant, bloquant ou transformant l'action.
Cette organisation rend visible le caractère médié du couplage perception–action. Reste à préciser comment ces médiations sont configurées dans le prompt.

2.2 Le prompt comme agrégat sémiotique

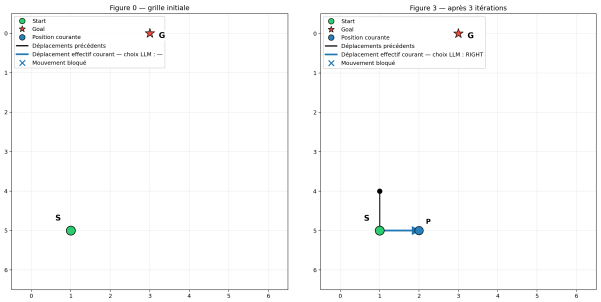
Dans le dispositif étudié ici, le prompt ne se réduit pas à une instruction isolée. Il constitue la forme textuelle sous laquelle plusieurs composantes du couplage perception-action sont rassemblées et rendues adressables au modèle. Il agrège ainsi le cadrage de la tâche, l'observation courante, l'historique disponible et les contraintes de réponse.
Cette précision évite une confusion de niveaux. L'observation, l'historique et les règles de réponse ne sont pas extérieurs au prompt : ils y sont intégrés. En revanche, le contrôleur qui interprète la sortie du modèle et l'environnement qui en détermine l'effet lui demeurent extérieurs. Le prompt ne se confond donc pas avec le dispositif, même s'il en concentre plusieurs médiations décisives.
Trois niveaux peuvent ici être distingués :

1. **Dans le prompt** : le cadrage de la tâche, l'observation courante, l'historique disponible et les contraintes de réponse.
2. **Hors du prompt, côté contrôleur** : l'interprétation de la sortie du modèle et sa conversion en action exécutable.
3. **Hors du prompt, côté environnement** : la validation, le blocage ou la transformation de l'action dans la grille.

Ce dispositif a été programmé sous la forme d'un module au sein de la plateforme *Hermes-HSF* qui sera prochainement distribué en open-source.

2.3 Mise en œuvre informatique

Cette mise en œuvre expérimentale repose sur une grille discrète partiellement inconnue, un signal local minimal, un espace d'actions borné et un grand modèle de langage inséré dans une boucle contrôlée perception-action. La plateforme *Hermes-HSF* assure la médiation entre l'environnement, la textualisation du signal, l'espace d'actions autorisé et l'exécution des actions produites par le modèle.
La grille comprend une position de départ, une position cible et, selon les cas, des contraintes spatiales limitant les déplacements possibles. À chaque itération, l'agent reçoit une observation minimale, notée `OBS`, qui ne fournit qu'une information locale sur son rapport immédiat à l'environnement. Il ne dispose donc d'aucune représentation globale de l'espace. L'enjeu n'est pas de résoudre une navigation sur carte connue, mais d'examiner la construction progressive de l'action sous contrainte.



État initial. État après trois itérations.

FIGURE 3 – État initial et état atteint après trois itérations dans l'environnement de grille.

TABLE 1 – Résumé des trois premières itérations.

Étape	Fig.	Action	Transition
Initial	0	–	–
1	1	`UP`	`(5,2)`→`(5,1)`
2	2	`DOWN`	`(5,1)`→`(5,0)`
3	3	`RIGHT`	`(5,0)`→`(5,0)`

L'observation est adressée au modèle par l'intermédiaire d'un prompt structuré. Le modèle ne produit pas directement une action exécutable. Il produit un signe d'action textuel. Un contrôleur externe en vérifie la conformité et le convertit en déplacement effectif dans la grille. L'environnement valide alors, bloque ou transforme ce déplacement selon sa structure propre.
Les figures 3 et le tableau 1 donnent un premier aperçu concret de cette chaîne. La figure montre l'état initial de la grille puis l'état atteint après trois itérations. Le tableau en détaille les actions choisies et les transitions correspondantes.
Le dispositif distingue trois niveaux d'analyse :

1. le monde n'est pas donné au modèle comme tel, mais sous la forme d'une textualisation contrainte d'un signal local ;
2. l'action du modèle ne s'exécute pas directement dans l'environnement, mais prend d'abord la forme d'un signe d'action ;
3. le contrôleur assure l'interface entre ce signe d'action et son exécution effective dans l'environnement, lequel conserve son extériorité propre en validant, bloquant ou transformant l'action proposée.

Nous comparons trois politiques d'action : *Baseline*, *No-Stay* et *Stay-Banned*. La *Baseline* sert ici de référence interne, non de comparaison avec un agent externe. Elles font varier, à modèle et environnement constants, la place de l'immobilité dans l'action et l'exécution.

Baseline Politique de référence dans laquelle l'ensemble complet des actions autorisées est proposé au modèle.

No-Stay Politique dans laquelle l'action `STAY` est explicitement retirée de l'ensemble des actions proposées au modèle.

Stay-Banned Politique dans laquelle l'immobilité n'est plus seulement absente des actions proposées au modèle, mais explicitement exclue au moment de l'exécution.

Dans un cadre d'interaction minimale, `STAY` ne constitue donc pas une action neutre. Elle peut signaler une inertie locale ou une difficulté à convertir le signal perçu en déplacement effectif. La comparaison de ces trois conditions permet d'isoler le rôle de l'immobilité dans la dynamique d'action du modèle.

2.4 Conditions d'inférence et choix des modèles

Les expériences sont conduites à paramètres de génération constants afin de préserver la comparabilité des résultats. Les écarts observés peuvent ainsi être rapportés au dispositif, au modèle et à la politique d'action, plutôt qu'à une variation des conditions de décodage. Les principaux paramètres sont résumés dans le tableau 2. Le choix de Llama3 et de Mistral permet de ne pas rapporter les résultats à une seule architecture. Tous deux sont instruction-tunés, disponibles dans Ollama et compatibles avec une mise en œuvre locale du protocole. Leur taille comparable, qui limite volontairement les effets liés à l'échelle, permet de concentrer l'analyse sur les effets du dispositif et des médiations plutôt que sur une comparaison de performance entre modèles de capacités très différentes.

TABLE 2 – Paramètres de génération et de contrôle utilisés pour les expériences.

Paramètre	Valeur	Justification
`temperature`	0.0	Supprime l'aléa de décodage et rend la comparaison entre politiques d'action plus directement interprétable.
`top_p`	1.0	N'introduit pas de restriction probabiliste supplémentaire et laisse agir le décodage déterministe induit par `temperature`=0.0.
`num_predict`	16	Offre une marge suffisante pour produire une sortie conforme sans tronquer artificiellement la génération.
`timeout_s`	180	Évite qu'un arrêt prématuré de l'inférence ne soit confondu avec un effet du dispositif.
`k_hist`	10	Maintient une fenêtre d'historique bornée et constante dans le prompt, compatible avec la logique d'interaction minimale.
`warmup`	1	Stabilise l'appel initial au modèle avant la collecte effective des résultats.

Afin de renforcer la reproductibilité du protocole, chaque exécution produit une trace structurée comprenant l'observation adressée au modèle, le prompt construit, la sortie brute, l'action parsée, l'éventuel recours au *fallback*, l'action effectivement exécutée, la position avant et après exécution, ainsi que l'état de succès de l'épisode. Les comparaisons sont effectuées à environnement, horizon maximal, espace d'actions et paramètres d'inférence constants. Les variations observées peuvent ainsi être rapportées aux médiations étudiées, modèle, politique d'action ou structuration du prompt, plutôt qu'à une variation implicite du dispositif expérimental. La section suivante introduit les métriques retenues pour caractériser le fonctionnement de cette architecture et les régimes d'action qu'elle produit.

3 Métriques de validation et d'analyse comportementale

L'évaluation du dispositif s'appuie sur deux familles de métriques complémentaires : des **métriques de contrôle**, qui garantissent l'interprétabilité des résultats, et des **métriques comportementales**, qui décrivent la manière dont une politique d'action se déploie effectivement dans l'environnement.

3.1 Métriques de contrôle

Les métriques de contrôle précèdent l'analyse proprement dite. Elles ne visent pas à caractériser directement la politique d'action. Elles servent à vérifier que les écarts observés entre modèles, politiques d'action ou variantes de prompt ne proviennent pas principalement d'un non-respect du format attendu ou d'un recours excessif au mécanisme de substitution. Elles portent notamment sur le respect du format strict des sorties, la conformité des épisodes et la fréquence de recours au *fallback*. Vérifiées en amont, elles ne sont pas développées davantage dans le corps de l'article, qui se concentre sur les métriques comportementales.

3.2 Métriques comportementales

Les métriques comportementales constituent le centre de l'analyse. Elles permettent de caractériser non seulement la réussite de la tâche, mais aussi le régime d'action produit sous signal minimal. Cinq métriques principales sont retenues.

1. **Le taux de succès** (r_{succ}) correspond à la proportion d'épisodes dans lesquels la cible est atteinte avant la borne maximale d'itérations :

$$r_{\text{succ}} = \frac{N_{\text{succ}}}{N_{\text{ep}}},$$

où N_{succ} désigne le nombre d'épisodes réussis et N_{ep} le nombre total d'épisodes.

2. **Le nombre moyen de pas** ($\bar{T}$) correspond au nombre moyen d'itérations effectivement exécutées par épisode :

$$\bar{T} = \frac{1}{N_{\text{ep}}} \sum_{i=1}^{N_{\text{ep}}} T_i,$$

où T_i désigne le nombre de pas exécutés dans l'épisode i.

3. **Le taux d'immobilité** (r_{stay}) est défini par

$$r_{\text{stay}} = \frac{N_{\text{STAY}}}{N_{\text{act}}},$$

où N_{STAY} désigne le nombre d'actions STAY effectivement exécutées et N_{act} le nombre total d'actions exécutées. Un taux élevé peut signaler une inertie locale, une prudence excessive ou une difficulté à convertir le signal minimal en déplacement effectif.

4. **L'entropie des actions** (H_{act}) est calculée à partir de la distribution empirique des actions exécutées :

$$H_{\text{act}} = -\sum_{a \in \mathcal{A}} p(a) \log p(a),$$

où $p(a)$ désigne la fréquence relative de l'action a.

5. **La couverture spatiale** (C_{sp}) mesure la proportion de positions effectivement visitées dans la grille :

$$C_{\text{sp}} = \frac{N_{\text{vis}}}{N_{\text{pos}}},$$

où N_{vis} désigne le nombre de positions distinctes visitées et N_{pos} le nombre total de positions accessibles. Lorsque cela s'avère utile, une couverture moyenne par épisode est également mobilisée :

$$\bar{C}_{\text{ep}} = \frac{1}{N_{\text{ep}}} \sum_{i=1}^{N_{\text{ep}}} C_{\text{sp}}^{(i)}.$$

Ces métriques remplissent des fonctions distinctes mais complémentaires : r_{succ} et $\bar{T}$ renseignent sur l'efficacité globale de la politique ; r_{stay} et H_{act} sur sa structure comportementale ; C_{sp} sur son extension effective dans l'environnement.

4 Validation empirique et comparaison des politiques d'action

Cette section poursuit un triple objectif. Elle établit d'abord l'exploitabilité empirique du dispositif, en montrant que les comportements observés ne procèdent pas principalement d'artefacts de format ou d'un recours au *fallback*. Elle compare ensuite, à prompt constant, les trois politiques d'action retenues, *Baseline*, *No-Stay* et *Stay-Banned*, sur les deux modèles considérés, Mistral et Llama3. Elle vérifie enfin que les contrastes observés en condition *Baseline* demeurent robustes à une variation limitée du tirage aléatoire du dataset. Il ne s'agit donc pas encore d'examiner l'effet d'une variation sémiotique du prompt, mais d'établir une première base de comparaison robuste dans le cadre de l'interaction minimale.

4.1 Contrôle expérimental du dispositif et comparaison avec une politique aléatoire

L'examen des différences comportementales suppose d'abord de vérifier que les résultats ne sont pas dominés par des artefacts de format ou par le recours au *fallback*. Nous ajoutons également une politique aléatoire non-LLM en condition de référence. Cette politique, notée *Random*, ne reçoit ni prompt, ni historique, ni interprétation de l'observation : elle sélectionne uniformément une action admissible. Elle ne constitue donc pas une politique comparable aux LLM, mais un repère minimal d'exploration non linguistique.

TABLE 3 – Politique aléatoire non-LLM et LLM en condition de référence, sur 100 épisodes et $T_{\max} = 100$.

Agent	Succès	Pas	STAY	H_{act}	MaxAct	StrictStep	Fallback
Random	0.35	84.36	0.199	1.575	0.264	1.000	0.000
Mistral	0.10	95.84	0.010	1.417	0.273	1.000	0.000
Llama3	0.12	94.32	0.000	1.364	0.300	1.000	0.000

Cette comparaison ne vise pas à faire du taux de succès le critère principal, mais à distinguer une exploration aléatoire d'un régime d'action médié. Si les LLM se confondaient avec *Random*, leurs distributions d'actions, leur recours à `STAY` et leur concentration seraient similaires. Les écarts observés permettent au contraire d'interpréter leurs sorties comme le produit d'un couplage spécifique entre modèle, prompt, format de réponse, contrôleur et environnement.
Le tableau 3 montre que la comparaison est exploitable : les trois agents atteignent un *strict step rate* parfait et aucun recours au *fallback*. Les écarts ne procèdent donc ni d'un défaut de format ni d'une substitution par le contrôleur. La politique aléatoire obtient un taux de succès supérieur aux deux modèles, ce qui confirme que la réussite brute ne constitue pas le critère principal du dispositif. En revanche, Mistral et Llama3 se distinguent de *Random* par un recours quasi nul à `STAY`, une entropie plus faible et une distribution d'actions plus concentrée. Les LLM ne se réduisent donc pas à une exploration aléatoire, même lorsqu'ils ne la surpassent pas en succès brut.
Ce résultat précise la portée méthodologique du protocole. Une lecture centrée sur le seul taux de succès ferait apparaître la politique aléatoire comme supérieure aux deux modèles. Mais ce serait transformer l'expérience en simple test de navigation, alors qu'elle vise à caractériser un régime d'action médié. La comparaison avec *Random* permet justement de dissocier performance brute et structure du comportement. Elle montre qu'une exploration aléatoire peut atteindre plus souvent la cible, tout en produisant un rapport distinct à l'immobilité, à l'entropie et à la concentration des actions. Les LLM, bien que moins performants dans cette condition, ne se confondent pas avec ce régime aléatoire. Ce déplacement est central : l'objet de l'analyse n'est pas la réussite isolée, mais la manière dont une sortie textuelle devient action au sein d'une chaîne de médiations comprenant prompt, format, contrôleur et environnement.

4.2 Comparaison comportementale des trois politiques d'action

Cette première vérification permet désormais d'examiner la dimension proprement comportementale des résultats. Le tableau 4 rassemble les valeurs agrégées sur 10 épisodes

par configuration. Deux niveaux de lecture doivent ici être distingués : le premier concerne l'articulation entre réussite et inertie ; le second porte sur la structure de l'exploration.

TABLE 4 – Comparaison des trois politiques d'action sur Mistral et Llama3, à prompt constant, sur 10 épisodes et $T_{\max} = 100$.

Modèle	Politique	Succès	Pas moy.	Stay	Entropie	Coverage
Mistral	Baseline	0.100	94.0	0.249	2.240	0.592
Mistral	No-Stay	0.300	79.4	0.000	1.947	0.735
Mistral	Stay-Banned	0.100	92.4	0.062	2.144	0.755
Llama3	Baseline	0.400	68.5	0.000	1.110	0.633
Llama3	No-Stay	0.400	68.4	0.000	1.660	0.612
Llama3	Stay-Banned	0.400	71.2	0.000	1.643	0.612

La relation entre réussite et inertie apparaît d'abord dans la figure 4.

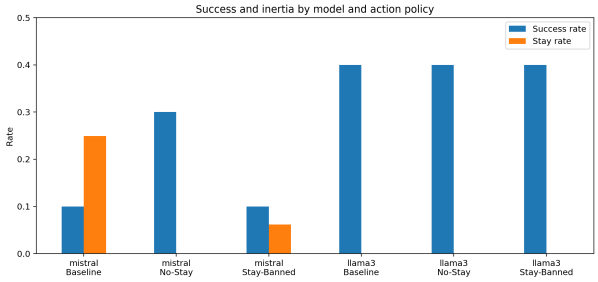


FIGURE 4 – Succès et inertie selon le modèle et la politique d'action. Pour le modèle Mistral, *No-Stay* réduit fortement l'immobilité et améliore la réussite. Pour le modèle Llama3, le taux de succès reste stable et l'immobilité est nulle dans les trois conditions.

En *Baseline*, une différence inter-modèles nette apparaît. Le modèle Llama3 obtient un taux de succès supérieur à celui du modèle Mistral tout en nécessitant moins de pas. L'écart porte également sur le rapport à l'immobilité : le modèle Mistral recourt fortement à `STAY`, alors que le modèle Llama3 ne l'utilise pas. Le tableau 4 en précise l'ampleur : le taux de succès de Llama3 atteint 0.4 contre 0.1 pour Mistral, tandis que le taux d'immobilité de Mistral s'établit à 0.249, contre 0 pour Llama3. La condition *Baseline* ne produit donc pas le même régime d'action selon le modèle considéré.

Pour le modèle Mistral, les politiques d'action modifient sensiblement l'articulation entre inertie et réussite. Le passage à *No-Stay* s'accompagne d'une baisse nette de l'immobilité et d'une hausse du succès. Le tableau 4 en confirme l'ampleur : le succès passe de 0.1 à 0.3 et le nombre moyen de pas de 94.0 à 79.4. La condition *Stay-Banned*, en revanche, réduit fortement l'immobilité sans améliorer le taux de succès. Les politiques d'action n'affectent donc pas toutes de la même manière la dynamique du modèle Mistral : certaines paraissent favoriser une exploration plus efficace, tandis que d'autres modifient le régime d'action sans gain net de réussite.

Les résultats obtenus avec le modèle Llama3 appellent une lecture distincte. Le taux de succès demeure stable dans les trois conditions, tout comme l'absence de `STAY`. À ce niveau, l'effet des politiques d'action ne se laisse donc pas saisir d'abord par l'inertie ou par la réussite. Il doit être recherché du côté de la structure de l'exploration.

La figure 5 permet alors d'examiner cette seconde dimension à partir de l'entropie des actions et de la couverture spatiale.

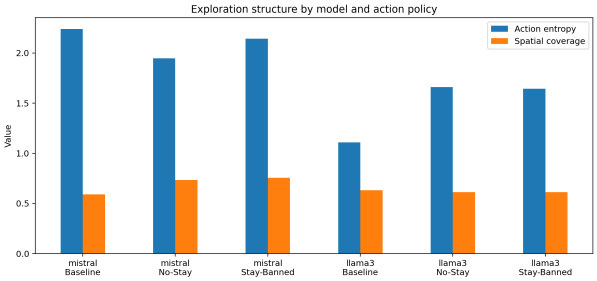


FIGURE 5 – Entropie des actions et couverture spatiale selon le modèle et la politique d'action. Les politiques d'action n'affectent pas les deux modèles de la même manière.

Les résultats obtenus avec le modèle Mistral montrent que les deux politiques modifiées produisent ici encore des effets distincts. *No-Stay* augmente nettement la couverture spatiale tout en améliorant la réussite. *Stay-Banned*, pour sa part, conduit à la couverture spatiale la plus élevée des trois conditions, mais sans gain sur le taux de succès. L'extension de l'exploration ne se traduit donc pas mécaniquement par une meilleure performance. Ce résultat suggère que la modification de la politique d'action n'agit pas seulement sur la quantité d'exploration, mais aussi sur sa forme.

Les résultats obtenus avec le modèle Llama3 font apparaître une dynamique différente. La couverture spatiale varie peu, mais l'entropie augmente nettement dans *No-Stay* et *Stay-Banned* par rapport à *Baseline*. Le tableau 4 en précise la mesure : l'entropie passe de 1.110 en *Baseline* à 1.660 et 1.643 dans les deux autres conditions. Autrement dit, les politiques d'action ne modifient pas sensiblement la réussite globale de Llama3, mais reconfigurent la distribution des actions. La condition de référence apparaît ainsi comme plus polarisée, tandis que les deux autres introduisent une plus grande diversité sans bénéfice net sur le succès.

Pris ensemble, le tableau 4 et les figures 4 et 5 font apparaître un résultat plus fort qu'une simple différence entre modèles ou entre politiques d'action prises séparément. Le même changement de politique d'action ne produit pas les mêmes effets selon le modèle considéré. Pour le modèle Mistral, les politiques d'action modifient à la fois l'inertie, la réussite et l'extension de l'exploration. Pour le modèle Llama3, elles affectent surtout la structure de la politique, avec une performance globale plus stable. Le résultat central est donc un effet de couplage entre architecture du modèle et structuration de l'espace d'action. Les comporte-

ments observés ne relèvent ni du seul modèle, ni de la seule politique d'action, mais de leur interaction dans le dispositif étudié.

Ces écarts peuvent être rapprochés, avec prudence, de certaines caractéristiques documentées des deux modèles considérés. Llama3.1 8B Instruct est présenté comme un modèle instruction-tuné optimisé pour des usages de dialogue multilingue, ce qui est compatible avec la très forte stabilité formelle observée ici ainsi qu'avec la relative constance de ses performances selon les politiques d'action [7]. Mistral 7B Instruct v0.3, également instruction-tuné, est présenté comme une version de Mistral-7B-v0.3 prenant en charge le tokenizer v3 et le *function calling* [8]. Dans notre protocole, ce modèle apparaît plus sensible aux modifications du régime d'action. Un tel rapprochement ne vaut toutefois pas explication causale. Les résultats présentés ici n'autorisent pas à attribuer les écarts observés à des propriétés intrinsèques des modèles seuls. Ils indiquent d'abord que ces écarts prennent sens au sein du couplage entre architecture du modèle et structuration du dispositif.

Du point de vue de l'interaction minimale, ces résultats vont dans le sens attendu. Le comportement observé ne prend pas la forme d'une navigation appuyée sur une représentation globale du monde, mais celle d'une orientation progressive sous signal pauvre. Du point de vue de la fabrique sémiotique hybride, ils apportent ainsi un premier appui empirique à l'idée selon laquelle l'action produite dépend de l'agencement des médiations qui définissent le régime d'exécution. Cette analyse établit ainsi une première base de comparaison expérimentale. Sur cette base, la section suivante se recentre sur la seule condition *Baseline*, afin d'examiner plus précisément les effets d'une variation sémiotique du prompt dans un dispositif d'interaction minimale.

4.3 Sensibilité des résultats au tirage aléatoire du dataset

Les résultats précédents reposent sur un premier tirage du dataset. Un test limité de sensibilité au tirage aléatoire des épisodes a donc été conduit afin d'en apprécier la robustesse. Il ne s'agissait pas de proposer une étude exhaustive de variabilité, mais de vérifier si les tendances observées en condition *Baseline* demeuraient lisibles lorsque seule la structure du dataset variait. Dans ce test complémentaire, seule la graine aléatoire de génération du dataset est modifiée ; les paramètres d'inférence du modèle demeurent inchangés. Pour une même graine, Mistral et Llama3 sont en outre évalués sur exactement le même jeu d'épisodes.

Trois graines aléatoires ont ainsi été considérées, avec 50 épisodes par graine et un horizon maximal de 100 itérations. Le tableau 5 fait apparaître une forte stabilité des profils observés. Quelques exemples suffisent à en mesurer la portée. Pour le modèle Mistral, le taux de succès demeure strictement stable à 22 % pour les trois tirages, tandis que le nombre moyen de pas varie seulement de 84.76 à 84.84. Pour le modèle Llama3, le taux de succès reste compris entre 28 % et 30 %, avec un nombre moyen de pas compris entre 73.86 et 75.18. Dans les deux cas, les autres indicateurs évoluent très peu : le taux d'immobilité demeure stable, l'entropie varie à peine et le *strict step rate* reste constant à 99.3 % pour Mistral et à 100 % pour Llama3.

Ces résultats indiquent que, dans la condition *Baseline*, les différences observées entre Mistral et Llama3 ne reposent pas sur un tirage particulier des épisodes. Elles apparaissent robustes à une variation limitée de la graine aléatoire du dataset. Ce point renforce l'interprétation proposée dans cette section : les contrastes entre modèles relèvent bien d'un effet de couplage entre architecture du modèle et structuration du dispositif, et non d'un artefact imputable à un seul jeu d'épisodes.

TABLE 5 – Sensibilité des résultats à la graine aléatoire du dataset en condition *Baseline*, sur 50 épisodes et $T_{\max} = 100$.

Modèle	Graine	Succès	Pas	Stay	Entr.	Cov.	Strict
Mistral	42	22.0	84.84	25.39	2.247	71.43	99.3
Mistral	43	22.0	84.84	25.46	2.247	71.43	99.3
Mistral	44	22.0	84.76	25.46	2.247	71.43	99.3
Mistral	moy.	22.0	84.81	25.44	2.247	71.43	99.3
Mistral	éc.-t.	0.0	0.05	0.04	0.000	0.00	0.0
Llama3	42	30.0	73.86	0.00	1.216	69.39	100.0
Llama3	43	28.0	75.14	0.00	1.212	69.39	100.0
Llama3	44	28.0	75.18	0.00	1.212	69.39	100.0
Llama3	moy.	28.7	74.73	0.00	1.213	69.39	100.0
Llama3	éc.-t.	1.15	0.75	0.00	0.002	0.00	0.0

5 Progression du prompt et reconfiguration du régime d'action

La section précédente a montré que, dans la condition *Baseline*, le comportement observé dépend déjà du couplage entre modèle et politique d'action. Il s'agit désormais d'examiner un autre niveau du dispositif, celui de la structuration du prompt. Dans cette perspective, seule la condition *Baseline* est retenue, et les médiations textuelles adressées au modèle sont enrichies progressivement.

L'enjeu n'est ni d'identifier la meilleure politique au sens d'une optimisation de la performance brute, ni de proposer un benchmark compétitif au sens classique. Le dispositif vaut ici comme instrument de discrimination conceptuelle : il s'agit de rendre empiriquement discernables les effets de la structuration du prompt sur le comportement observé.

Trois versions du prompt sont considérées. La première, *P0*, correspond à un prompt minimal sans historique ni indication supplémentaire. La deuxième, *P1*, ajoute un historique court des pas précédents. La troisième, *P2*, reprend ce même historique et lui adjoint une indication interprétative explicite invitant le modèle à l'utiliser pour éviter des déplacements erratiques. Les trois versions sont évaluées sur le même ensemble de 100 épisodes, de manière à isoler l'effet propre de cet enrichissement progressif.

5.1 Du prompt minimal à l'historique : sortie d'une politique dégénérée

Le tableau 6 rassemble les résultats obtenus sur 100 épisodes pour les six configurations considérées.

TABLE 6 – Effets de l'enrichissement progressif du prompt sur 100 épisodes et $T_{\max} = 100$.

Modèle	Prompt	Succès	Pas	Stay	Entr.	Cov.
Mistral	P0 min.	0.000	100.00	1.000	0.000	0.469
Mistral	P1 hist.	0.220	86.04	0.254	2.246	0.714
Mistral	P2 hist.+ind.	0.020	99.26	0.655	1.600	0.878
Llama3	P0 min.	0.000	100.00	0.000	0.000	0.510
Llama3	P1 hist.	0.250	78.45	0.000	1.227	0.735
Llama3	P2 hist.+ind.	0.240	80.20	0.000	1.747	0.653

La figure 6 fait apparaître l'articulation entre taux de succès et taux d'immobilité selon le niveau d'enrichissement du prompt.

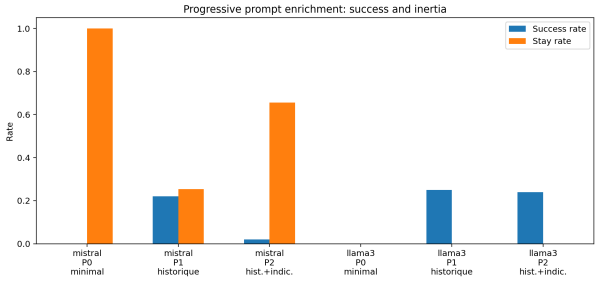


FIGURE 6 – Succès et inertie selon le modèle et le niveau d'enrichissement du prompt. Le passage de *P0* à *P1* fait réapparaître une politique d'action efficace pour les deux modèles.

Le prompt minimal ne soutient, dans ce dispositif, aucune politique d'action viable. Pour les deux modèles, le taux de succès est nul, l'horizon maximal est systématiquement atteint et l'entropie des actions demeure nulle. La dégénérescence observée ne prend toutefois pas la même forme selon le modèle considéré. Le modèle Mistral se fige dans l'immobilité, avec un taux de `STAY` égal à 1. Le modèle Llama3 n'utilise pas `STAY`, mais reste néanmoins enfermé dans une politique entièrement polarisée, sans diversification effective de l'action.

Le passage à *P1* transforme fortement cette situation. Le succès réapparaît pour les deux modèles, le nombre moyen de pas diminue, l'entropie augmente et la couverture spatiale progresse nettement. Pour Mistral, le taux de succès passe de 0.000 à 0.220 et la couverture de 0.469 à 0.714. Pour Llama3, le succès passe de 0.000 à 0.250 et la couverture de 0.510 à 0.735. L'historique constitue ainsi une médiation textuelle décisive : sans lui, l'action demeure incapable de s'orienter efficacement dans le dispositif.

5.2 L'indication interprétative : une reconfiguration non uniforme

La figure 7 permet d'examiner l'effet de l'ajout d'une indication interprétative à partir de l'entropie des actions et de la couverture spatiale.

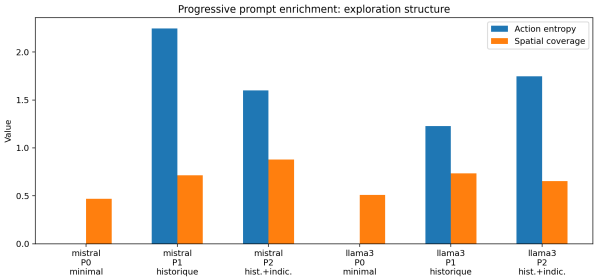


FIGURE 7 – Entropie des actions et couverture spatiale selon le modèle et le niveau d'enrichissement du prompt. Le passage à *P2* n'affecte pas les deux modèles de la même manière.

L'ajout d'une indication interprétative ne produit pas un effet uniforme. Pour le modèle Mistral, le passage de *P1* à *P2* s'accompagne d'une chute du taux de succès, qui passe de 0.220 à 0.020, d'un retour vers l'horizon maximal, d'une forte hausse de l'immobilité, de 0.254 à 0.655, et d'une baisse de l'entropie. La couverture spatiale augmente pourtant nettement, de 0.714 à 0.878. L'effet observé ne correspond donc pas à une amélioration linéaire de la politique. Il prend la forme d'une dissociation entre capacité de convergence et extension de l'exploration.

Pour le modèle Llama3, l'effet est différent. Le succès demeure presque stable, passant de 0.250 à 0.240, sans apparition d'immobilité. En revanche, l'entropie augmente fortement, de 1.227 à 1.747, tandis que la couverture spatiale diminue, de 0.735 à 0.653. L'indication interprétative ne désorganise donc pas la politique. Elle en reconfigure la distribution locale, sans bénéfice net sur la réussite.

Les métriques de contrôle permettent d'exclure l'hypothèse d'un effondrement formel du dispositif. La figure 8 en présente l'état dans les six configurations.

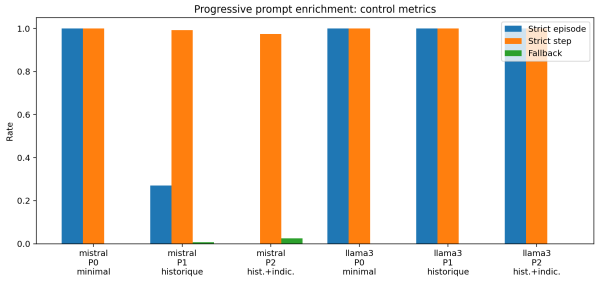


FIGURE 8 – Métriques de contrôle selon le modèle et le niveau d'enrichissement du prompt. Les sorties de Llama3 demeurent parfaitement conformes dans les trois conditions, tandis que Mistral devient plus fragile en *P2*.

Les sorties de Llama3 demeurent parfaitement conformes dans les trois conditions. Mistral devient, en revanche, plus fragile en *P2*, avec une baisse du *strict step rate* et une légère augmentation du *fallback*. Cette fragilité ne suffit pas à expliquer à elle seule la transformation observée du comportement, mais elle indique que l'ajout de l'indication accroît aussi, pour ce modèle, la vulnérabilité formelle de la boucle.

Trois résultats ressortent du tableau 6 et des figures 6, 7 et 8. Le prompt minimal ne suffit pas, dans ce dispositif, à soutenir une politique d'action viable. L'ajout d'un historique réintroduit, pour les deux modèles, une capacité d'orientation, d'exploration et de réussite. L'ajout d'une indication interprétative ne produit pas, enfin, un gain uniforme, mais une reconfiguration différentielle du régime d'action selon le modèle considéré.

Ce dernier point est décisif. Le même enrichissement du prompt n'a pas les mêmes effets selon l'architecture du modèle. Pour Mistral, il tend à accroître l'inertie et à désorganiser la réussite tout en augmentant fortement la couverture spatiale. Pour Llama3, il maintient une réussite globalement stable, sans immobilité, mais reconfigure la distribution locale des actions. Le résultat central n'est donc pas seulement qu'un prompt agit, ni même qu'un historique agit, mais que l'enrichissement progressif du prompt révèle un effet de couplage entre architecture du modèle et structuration des médiations textuelles.

Du point de vue de l'interaction minimale, ces résultats confirment que l'action observée ne procède pas d'une représentation globale du monde, mais d'une orientation locale sous signal pauvre, rendue possible ou transformée par la manière dont le dispositif textualise le passé immédiat. Du point de vue de la fabrique sémiotique hybride, ils montrent que le prompt ne constitue pas un simple support d'instruction, mais un opérateur de médiation dont l'enrichissement progressif infléchit directement le régime d'action. L'action effective ne résulte ainsi ni du seul modèle, ni du seul environnement, ni d'un texte pris isolément. Elle dépend d'un agencement de médiations dont la structure devient ici empiriquement observable. Ces résultats ne permettent évidemment pas de valider à eux seuls l'ensemble du concept de fabrique sémiotique hybride. Ils apportent toutefois un premier appui empirique à l'hypothèse selon laquelle l'action observée dépend de l'agencement des médiations qui structurent le dispositif, et non du seul modèle pris isolément.

6 Discussion

Les résultats apportent un premier appui empirique à l'hypothèse directrice de cet article. Ils montrent que le cadre d'interaction minimale rend observables les médiations nécessaires à la production d'une action par un grand modèle de langage engagé dans une boucle perception–action.

Ils éclairent en particulier le statut du prompt. Les variations de sa structuration modifient le régime d'action observé, ce qui confirme qu'il ne se réduit pas à un simple habillage instructionnel. Plus largement, le comportement produit ne peut être rapporté ni au seul modèle, ni au seul prompt, ni au seul environnement. Il doit être rapporté à leur agencement.

Une telle interprétation demande toutefois à être confirmée. Elle prolonge, sur le plan épistémologique, le travail engagé ici autour du concept de fabrique sémiotique hybride. L'action n'y apparaît plus comme l'effet d'une boucle de contrôle homogène. Elle apparaît comme le produit de relations distribuées entre modèle, textualisation, contrôleur et environnement. En ce sens, elle peut être rapprochée, avec prudence, du déplacement proposé par Edwin Hutchins, lorsqu'il analyse la cognition au niveau des relations entre agents, artefacts et milieu [5].

Les expériences demeurent limitées en taille, en nombre de modèles et en variété de conditions. Elles suggèrent néanmoins que l'étude des grands modèles de langage orientés vers l'action gagne à être conduite moins en termes de performance isolée qu'en termes d'architecture de médiations.

Conclusion

Cet article propose une première mise à l'épreuve expérimentale de la fabrique sémiotique hybride dans un cadre d'interaction minimale. La plateforme *Hermes-HSF* insère un grand modèle de langage dans une boucle contrôlée perception–action, au sein d'un environnement discret, sous signal local pauvre et sous contraintes explicites d'exécution.

Les résultats montrent que l'action observée ne relève ni du seul modèle, ni du seul prompt, ni du seul environnement. Elle dépend d'un agencement de médiations associant signal, textualisation, contrôleur d'exécution et validation environnementale. Les comparaisons entre politiques d'action, les variations du prompt et le test limité de robustesse convergent vers un même constat : le comportement produit relève d'un couplage entre architecture du modèle et structuration du dispositif.

L'approche proposée relève moins d'un benchmark que d'une plateforme d'analyse des médiations permettant d'étudier les grands modèles de langage à partir des conditions qui rendent possible leur insertion dans des boucles perception–action. L'élargissement des modèles, des signaux et des textualisations permettra d'en éprouver la por-

tée.

Références

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, et al. Do as i can, not as i say : Grounding language in robotic affordances. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318, 2023.

[2] Malika Auvray, Charles Lenay, and John Stewart. Perceptual interactions in a minimalist virtual environment. *New Ideas in Psychology*, 27(1) :32–47, 2009.

[3] Malika Auvray and Marieke Rohde. Perceptual crossing : The simplest online paradigm. *Frontiers in Human Neuroscience*, 6 :181, 2012.

[4] Christophe Denis. Vers une coconstruction harmonieuse du sens entre langage humain et machinique au sein de la fabrique sémiotique hybride. *Socio-anthropologie*, 52 :37–50, 2025.

[5] Edwin Hutchins. *Cognition in the Wild*. MIT Press, 1995.

[6] Charles Lenay and John Stewart. Minimalist approach to perceptual interactions. *Frontiers in Human Neuroscience*, 6 :98, 2012.

[7] Meta. meta-llama/llama-3.1-8b-instruct. Hugging Face model card, 2024. Consulté en ligne.

[8] Mistral AI. mistralai/mistral-7b-instruct-v0.3. Hugging Face model card, 2024. Consulté en ligne.

[9] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct : Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

JK-Means : LLM-as-a-Judge au Service du Clustering Intentionnel

Arnaud Deleruyelle[1], Steve Bellart[1], Jérôme Mollier-Pierret[1]

[1] Talan – Centre de Recherche et d'Innovation, France

Résumé

Nous proposons JK-Means, un algorithme de clustering guidé par l'intention utilisateur en langage naturel. Notre contribution principale repose sur un mécanisme d'évaluation LLM-as-a-Judge en temps constant qui valide la cohérence des clusters et déclenche des opérations de split/-merge adaptatives. Évaluée sur trois jeux de données de référence, notre approche atteint une haute précision et un fort alignement sémantique avec les classes d'origine, démontrant sa capacité à découvrir des clusters pertinents et conformes aux attentes de l'utilisateur.

Mots-clés

Clustering, LLM as a Judge, Intention utilisateur, Embeddings sémantiques.

Abstract

We propose JK-Means, a clustering algorithm guided by user intent expressed in natural language. Our main contribution relies on a constant-time LLM-as-a-Judge evaluation mechanism that validates cluster coherence and triggers adaptive split/merge operations. Evaluated on three benchmark datasets, our approach achieves high accuracy and strong semantic alignment with the original classes, demonstrating its ability to discover relevant clusters that conform to user expectations.

Keywords

Clustering, LLM as a Judge, User intent, Semantic embeddings.

1 Introduction

Si le clustering permet d'organiser efficacement des données non annotées, son application se heurte souvent à une limite majeure : la déconnexion entre la formation des clusters et la volonté de l'utilisateur. Les algorithmes de type K-Means, largement plébiscités pour leur rapidité computationnelle, illustrent bien ce problème : leur logique purement géométrique et leur sensibilité à l'initialisation les empêchent d'intégrer l'expertise métier. Or, dans de nombreux contextes applicatifs (analyse de tickets de support, classification de documents, veille documentaire), l'utilisateur possède une intention spécifique sur la manière de catégoriser les données, une connaissance que les méthodes classiques ne permettent pas d'exploiter.

Face à ces limites, l'émergence des Grands Modèles de Langage (LLM) offre de nouvelles opportunités pour le clustering de textes. Ces modèles possèdent des capacités de compréhension sémantique riches et peuvent générer des descriptions textuelles interprétables. De plus, les embeddings modernes capturent des relations sémantiques complexes dans des espaces vectoriels de haute dimension. Cependant, les approches existantes utilisant les LLM pour le clustering se limitent souvent à la génération de labels *a posteriori* ou à la classification directe, sans véritablement guider la formation des clusters par l'intention de l'utilisateur.

Nous proposons **JK-Means**, un algorithme de clustering non supervisé guidé par l'intention utilisateur exprimée en langage naturel. Nos contributions s'articulent autour de trois mécanismes clés qui redéfinissent les étapes standards de K-LLMeans :

Un cadre de clustering intentionnel continu. L'intention de l'utilisateur guide désormais toutes les étapes de l'algorithme. Par exemple, pour l'intention « identifier les problèmes techniques SAP », l'algorithme génère automatiquement des centroïdes textuels spécifiques (« Erreurs de connexion à la base de données », « Problèmes d'autorisation », etc.), les raffine à chaque itération selon les documents assignés, et évalue leur pertinence par rapport au besoin métier. Cette approche transforme le clustering d'une tâche purement géométrique en une tâche sémantique ciblée, tout en restant non supervisée.

Une initialisation gloutonne maximisant la diversité. Pour contrer la sensibilité aux conditions initiales, nous proposons une heuristique de sélection des centroïdes textuels de départ. Afin d'éviter l'apparition de clusters "trop génériques" (ex : « Problèmes système ») qui absorbent trop de documents et créent des minima locaux, notre méthode sélectionne itérativement les titres les plus éloignés sémantiquement. Cela garantit un point de départ spécifique et bien segmenté.

Une évaluation LLM-as-a-Judge frugale et ciblée. Pour valider la cohérence des clusters sans faire exploser les coûts liés aux appels API des LLM, nous introduisons un mécanisme d'évaluation suivant : Plutôt que de faire analyser l'intégralité du jeu de données par le modèle, l'algorithme se concentre uniquement sur les cas les plus incertains : les documents les plus éloignés du centroïde. L'idée sous-jacente de cette approche étant que si ces documents périphériques sont jugés cohérents avec le titre du cluster, alors le cœur du cluster l'est également. Dans le cas contraire, un mécanisme de séparation (split) génère deux nouveaux titres plus spécifiques. Cette stratégie permet d'échapper aux minima locaux et d'affiner la catégorisation, tout en réduisant drastiquement les coûts computa-

tionnels et financiers.

Le reste de l'article est organisé comme suit : la section 2 présente l'état de l'art sur le clustering et les approches basées sur les LLM. La section 3 détaille notre méthode JK-Means. La section 4 présente le protocole expérimental et la section 5 se concentre sur les résultats obtenus. Enfin, la Section 6 conclut et discute des perspectives.

2 Travaux Antérieurs

2.1 Clustering Non Supervisé Classique

L'algorithme K-Means [10] reste une référence pour le clustering non supervisé grâce à sa simplicité et son efficacité. Il alterne entre assignation des points au centroïde le plus proche et mise à jour des centroïdes comme moyenne des points assignés. Cependant, K-Means souffre de limitations bien documentées : sensibilité à l'initialisation aléatoire, nécessité de fixer le nombre de clusters K a priori, et tendance à converger vers des minima locaux.

K-Means++ [2] améliore l'initialisation en sélectionnant les centroïdes initiaux de manière probabiliste. L'algorithme fonctionne comme suit : (1) le premier centroïde est choisi uniformément au hasard parmi les points de données, (2) pour chaque point x, on calcule la distance $D(x)$ au centroïde le plus proche déjà sélectionné, (3) le prochain centroïde est choisi avec une probabilité proportionnelle à $D(x)^2$, favorisant ainsi les points éloignés des centroïdes existants. Cette approche garantit que les centroïdes initiaux sont bien espacés, réduisant le risque de convergence vers des minima locaux et améliorant la qualité finale du clustering.

Les approches basées sur la densité comme DBSCAN [5] et HDBSCAN détectent automatiquement le nombre de clusters et gèrent les outliers, mais leur complexité computationnelle limite leur scalabilité. Le clustering hiérarchique offre une flexibilité via les dendrogrammes, mais reste difficile à interpréter pour de grandes quantités de données.

Dans le contexte du clustering de textes, l'utilisation d'embeddings sémantiques (Word2Vec [11], GloVe [13], BERT [3]) a significativement amélioré la qualité des clusters en capturant les relations sémantiques. Sentence-BERT [15] génère des représentations vectorielles de phrases entières, permettant un clustering plus cohérent. Les modèles d'embedding modernes comme text-embedding-3-large offrent des représentations de haute dimension (3072 dimensions) pouvant capturer des nuances sémantiques. Des travaux récents [12] ont démontré que les embeddings issus de LLM pré-entraînés (comme all-MiniLM-L6-v2 [19]) produisent des clusters significativement plus interprétables par des humains que les méthodes traditionnelles (doc2vec, LDA) sur des textes courts, avec une amélioration de 40% de la cohérence perçue. Cependant, ces approches restent non supervisées au sens strict : elles ne permettent pas d'intégrer l'expertise ou l'intention de l'utilisateur de manière structurée.

2.2 LLM pour le Clustering : K-LLMeans et NILC

Les LLM ont ouvert de nouvelles perspectives pour le clustering de textes. Plusieurs travaux récents explorent leur utilisation pour générer des labels de clusters *a posteriori*, améliorant l'interprétabilité sans modifier la formation des clusters [21]. TopicGPT [14] utilise des prompts structurés pour générer des descriptions hiérarchiques de topics, facilitant l'interprétation des résultats. Viswanathan et al. [17] proposent d'utiliser les LLM pour augmenter les représentations de documents et générer des contraintes par paires, permettant un clustering semi-supervisé efficace avec un nombre réduit d'annotations manuelles.

K-LLMeans. K-LLMeans [4] introduit l'utilisation de résumés textuels comme centroïdes, offrant une approche interprétable pour le clustering de textes. L'algorithme présente de nombreux avantages : (1) **Coûts faibles** car il repose essentiellement sur des embeddings pour l'assignation des documents, avec seulement K appels LLM par étape de résumé pour la mise à jour des centroïdes. (2) **Scalabilité** préservée grâce à l'utilisation d'embeddings modernes efficaces. (3) **Interprétabilité** améliorée car les centroïdes textuels sont directement lisibles et actionnables. (4) **Robustesse** grâce aux connaissances sémantiques apportées par les LLM.

ClusterLLM. Dans une approche complémentaire, ClusterLLM [21] exploite les LLM pour guider le fine-tuning d'embedders pré-entraînés. Le LLM est interrogé sur des triplets ⟨anchor, choice_1, choice_2⟩ pour déterminer quelle option est la plus proche de la référence selon une perspective donnée (ex : "Select based on topic"). Un échantillonnage entropique sélectionne les triplets les plus informatifs, permettant de réduire les coûts d'API tout en améliorant les performances sur 14 datasets testés. Les prédictions du LLM sont utilisées pour fine-tuner l'embedder via une loss contrastive, produisant un espace de représentation mieux aligné sur la perspective souhaitée.

NILC. Dans le cadre de NILC [18], l'approche est plus coûteuse mais offre une précision accrue. NILC utilise les LLM comme juges ("LLM as a Judge" [22]) pour évaluer la cohérence des clusters et guider leur formation. Cette approche pilote la procédure d'assignation via des Modèles de Langage, permettant une évaluation sémantique fine de la qualité des clusters. Le paradigme "LLM as a Judge" démontre l'efficacité de l'utilisation des LLM pour des tâches d'évaluation et de validation, au prix d'un nombre d'appels LLM plus élevé ($O(n^2)$ dans certaines configurations).

Notre approche se distingue par son caractère frugal : elle combine l'intégration de l'intention utilisateur en langage naturel avec une évaluation stratégiquement ciblée via le paradigme "LLM as a Judge", maintenant un coût constant de K appels LLM par itération. Là où NILC déploie le paradigme "LLM as a Judge" de manière exhaustive avec une complexité en $O(n^2)$, et où ClusterLLM nécessite un fine-tuning coûteux de l'embedder sur de nombreux triplets, notre méthode réalise une évaluation intentionnelle sélective en se concentrant uniquement sur les éléments ex-

trêmes de chaque cluster. Cette stratégie frugale maintient un coût constant indépendant de n, permettant de détecter efficacement les incohérences sémantiques et de déclencher, si nécessaire, un mécanisme de split de cluster. Elle corrige ainsi la trajectoire des centroïdes en temps réel, évitant la convergence vers des structures sémantiquement non pertinentes, tout en préservant le caractère non supervisé du clustering. Notre approche opère directement dans l'espace sémantique des résumés générés, offrant flexibilité et agnosticisme par rapport au choix de l'embedder, sans nécessiter de phase d'entraînement préalable.

Initialisation. Notre heuristique gloutonne s'inspire de K-Means++ en maximisant l'espacement des centroïdes initiaux, mais opère dans l'espace sémantique des titres générés par le LLM plutôt que dans l'espace des données. Nous remplaçons la sélection probabiliste par une sélection basée sur les relations sémantiques : à chaque étape, nous choisissons le titre ayant la plus faible similarité cosinus avec les titres déjà sélectionnés, garantissant une diversité sémantique maximale.

3 Notre contribution : JK-Means

3.1 Préliminaires : K-Means pour le Clustering de Textes

Soit un corpus de n documents textuels $\mathcal{D} = \{d_1, \ldots, d_n\}$. Chaque document d_i est représenté par un vecteur d'embedding $\mathbf{x}_i \in \mathbb{R}^d$ tel que :

$$\mathbf{x}_i = f_{emb}(d_i) \tag{1}$$

Nous supposons que tous les embeddings sont normalisés ($\|\mathbf{x}_i\| = 1$), rendant équivalentes la distance euclidienne et la similarité cosinus. L'objectif de K-Means est de partitionner ces n embeddings en k clusters, minimisant la variance intra-cluster :

$$\arg \min_{C_1,\ldots,C_k} \sum_{j=1}^{k} \sum_{i \in [C_j]} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \tag{2}$$

où C_j dénote l'ensemble des embeddings assignés au cluster j, $[C_j] = \{i | \mathbf{x}_i \in C_j\}$ dénote l'ensemble des indices assignés au cluster j, et $\boldsymbol{\mu}_j$ est le centroïde du cluster, calculé comme la moyenne des embeddings assignés :

$$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{i \in [C_j]} \mathbf{x}_i \tag{3}$$

L'algorithme K-Means assigne itérativement chaque embedding $\mathbf{x}_i$ au centroïde le plus proche et met à jour les centroïdes jusqu'à convergence après T itérations. Cependant, K-Means souffre de limitations bien documentées : sensibilité à l'initialisation et tendance à converger vers des minima locaux en raison de la nature non-convexe de sa fonction objectif [10, 9].

3.2 Préliminaires : K-LLMeans

Pour améliorer K-Means pour le clustering de textes, K-LLMmeans [4] introduit le concept de **summary-as-centroid** : périodiquement, les centroïdes numériques sont remplacés par des résumés textuels générés par un LLM, puis ré-encodés dans l'espace d'embedding. Formellement, la distinction principale réside dans le mécanisme de mise à jour des centroïdes : toutes les l itérations, la mise à jour standard de l'équation (3) est remplacée par :

$$\boldsymbol{\mu}_j = f_{\text{emb}}\left(f_{\text{LLM}}(\mathcal{P}_j)\right) \tag{4}$$

où :

- f_{emb} est la fonction d'embedding (ex : text-embedding-3-small, e5-large) ;
- f_{LLM} est la fonction de génération du LLM (ex : GPT-4o, Claude-3.7) ;
- $\mathcal{P}_j$ = Prompt $\left(\mathcal{I}, \{d_{z_i} \mid z_i \sim [C_j]\}_{i=1}^{m_j}\right)$ est le prompt construit à partir de :
  - $\mathcal{I}$: l'instruction (ex : *"The following is a cluster of online banking questions. Write a single question that represents the cluster concisely."*) ;
  - $\{d_{z_i}\}_{i=1}^{m_j}$: un échantillon de m_j documents du cluster C_j, où $z_i \sim [C_j]$ désigne un indice échantillonné sans remplacement et m est le nombre maximal de documents par prompt (typiquement $m = 10$ en mode *few-shot*, ou $m = |C_j|$ pour utiliser tous les documents du cluster).

Entre deux étapes de résumé, K-LLMmeans effectue des itérations standards de K-Means (assignation + mise à jour numérique des centroïdes), préservant ainsi l'objectif classique de minimisation de la variance intra-cluster. Le prompt utilisé pour la génération de résumés est donné en Annexe B.2.

Plutôt que de fournir tous les documents du cluster en entrée, le LLM traite un échantillon représentatif afin de fortement limiter le coût et préserver la taille du contexte en entrée au LLM.

Dans K-LLMmeans [4], les auteurs proposent d'utiliser un **échantillonnage K-means++** des embeddings du cluster pour sélectionner les m documents représentatifs à fournir au LLM. Cette approche maximise la distance entre les documents échantillonnés, pour diversifier l'aperçu du contenu sémantique du cluster. Les auteurs comparent empiriquement cette option avec plusieurs alternatives testées sur les embeddings text-embedding-3-small, notamment la sélection des m documents les plus proches du centroïde.

Notre approche diffère de K-LLMmeans en sélectionnant cette seconde stratégie, choix motivé par plusieurs considérations. Tout d'abord les **écarts de performance marginaux** entre les stratégies d'échantillonnage restent faibles (0.3–0.7% ACC sur CLINC/GoEmo). Ensuite, contrairement à K-LLMmeans qui utilise text-embedding-3-small, nous utilisons text-embedding-3-large et l'augmentation de la dimensionnalité renforce le phénomène de **concentration des distances** [1] ce qui va encore réduire les écart entre ces stratégies. Enfin, la sélection des m plus proches nécessite uniquement un tri par distance (complexité $O(n \log m)$), alors que K-means++ requiert un échantillonnage itératif (complexité $O(m \cdot n)$).

3.3 Architecture Générale de JK-Means

JK-Means étend K-LLMeans en intégrant l'intention utilisateur comme une variable au coeur de nos templates de prompts, que cela soit durant les itérations, B.2 mais aussi dès l'initialisation B.1 et également durant notre étape additionnelle d'évaluation ciblée des clusters B.3.
La Figure 1 présente l'architecture complète de notre approche.
Nous modifions l'initialisation aléatoire de K-Means ou probabiliste de K-Means++ pour un algorithme glouton afin de commencer avec des centroides cohérents avec l'intention de l'utilisateur. Ensuite, nous reprenons les mêmes première étapes de K-LLMeans dans notre boucle principale (mais avec notre propre prompt de regénération de titre avec une focalisation sur l'intention utilisateur B.2) suivi d'un mécanisme d'évaluation qui, en fonction de son résultat, va enclencher soit une séparation du cluster (split) si celui-ci est jugé incohérent, soit une sauvegarde dynamique si le cluster est validé. Enfin, un mécanisme de fusion permet de pallier les clusters de petites tailles car ils sont responsables d'optimums locaux.
Les sous-sections suivantes détaillent l'intégration de l'intention utilisateur dans notre approche lors de l'initialisation gloutonne (Section 3.4), dans la boucle principal (Section 3.5) ainsi que dans l'évaluation des clusters et les mécanismes de split/merge des clusters (Section 3.6) avant de finir sur les optimisations de coûts en appel LLM (Section 3.7).

3.4 Heuristique Greedy pour l'initialisation des titres

L'objectif de cette phase est d'éviter la création de **clusters attracteurs dominants** : une mauvaise initialisation qui génère un titre trop générique (ex : "Problème système") risque d'attirer la majeure partie des documents, piégeant l'algorithme dans un minimum local. Nous voulons obtenir des titres spécifiques qui couvrent le spectre de l'intention utilisateur (voir Annexe B, Prompt B.1) sans être génériques.
Le LLM génère d'abord N_{cand} titres candidats (typiquement $N_{cand} = 50$) basés sur l'intention utilisateur. Nous devons ensuite sélectionner K titres parmi ces candidats pour initialiser les clusters. Nous nous basons sur la similarité cosinus pour identifier les titres spécifiques. L'idée est qu'un titre ayant une faible similarité cosinus avec de nombreux autres doit avoir une faible généricité. Notre heuristique gloutonne fonctionne comme suit :
Étape 1 : Sélection du premier titre. Nous sélectionnons le titre t_1^* ayant la plus faible similarité cosinus moyenne avec tous les autres candidats :

$$t_1^* = \arg\min_{t \in \mathcal{T}_{cand}} \frac{1}{N_{cand} - 1} \sum_{t' \in \mathcal{T}_{cand} \setminus \{t\}} \cos(f_{emb}(t), f_{emb}(t')) \tag{5}$$

Étape 2 : Sélection gloutonne itérative. Une fois un titre sélectionné et ajouté au pool $\mathcal{T}_{sel}$, nous ajoutons itérativement les nouveaux titres un par un en choisissant celui qui minimise la similarité cosinus moyenne avec les titres déjà sélectionnés :

$$t_{i+1}^* = \arg\min_{t \in \mathcal{T}_{cand} \setminus \mathcal{T}_{sel}} \frac{1}{|\mathcal{T}_{sel}|} \sum_{t' \in \mathcal{T}_{sel}} \cos(f_{emb}(t), f_{emb}(t')) \tag{6}$$

Cette procédure garantit que chaque nouveau titre apporte une diversité sémantique maximale, forçant les clusters initiaux à être spécifiques et bien séparés. Les titres sélectionnés sont ensuite vectorisés pour former les centroïdes initiaux : $\mu_k = f_{emb}(t_k^*)$.

3.5 La régénération de titre avec intention

L'intention utilisateur guide l'ensemble du processus de clustering à travers les prompts LLM à plusieurs niveaux dont celui de la boucle principale visant à regénérer de nouveau titre de cluster. À chaque itération, lors du raffinement des titres, l'intention est intégrée dans le prompt pour générer des titres cohérents avec l'objectif métier (voir Annexe B, Prompt B.2). Cette intégration garantit que les titres générés restent alignés avec l'objectif métier tout au long des itérations.
De plus, dans la phase d'évaluation LLM as a Judge, l'intention se retrouve également dans le prompt d'évaluation (voir Annexe B, Prompt B.3) puisque le LLM évalue si un document appartient à un cluster en tenant compte de l'intention globale. Cette évaluation contextualisée garantit que les assignations respectent l'objectif métier et ne se basent pas uniquement sur la similarité sémantique brute de l'embedder.

3.6 Evaluation LLM as a Judge des clusters

À chaque itération, après l'assignation des documents aux clusters, nous évaluons la cohérence de chaque cluster pour identifier ceux nécessitant une subdivision (split) ou une fusion (merge). Cette évaluation repose sur le paradigme "LLM as a Judge" qui permet de détecter les clusters hétérogènes tout en maintenant un coût computationnel constant.

3.6.1 Évaluation LLM as a Judge

Pour chaque cluster C_k, nous évaluons la cohérence des N_{worst} documents les plus éloignés du centroïde (ceux avec les scores d'affinité les plus faibles). Soit D_{worst}^k l'ensemble de ces documents :

$$D_{worst}^k = \{d_i \in C_k \mid \text{affinity}(d_i, c_k) \in \text{Bottom}_{N_{worst}}(C_k)\} \tag{7}$$

Pour chaque document $d_i \in D_{worst}^k$, le LLM évalue si le document appartient bien au cluster en posant la question "Est-ce que d_i appartient à t_k ?" (voir Annexe B, Prompt B.3). Le score de cohérence du cluster est calculé comme le pourcentage de documents validés par le LLM parmi les N_{worst} documents évalués.
Si les documents les plus éloignés du centroïde (les plus difficiles à classer) sont validés par le LLM, alors nous considérons le cluster comme pertinent. Cette approche garantit un temps constant d'évaluation ($O(N_{worst})$ au lieu de

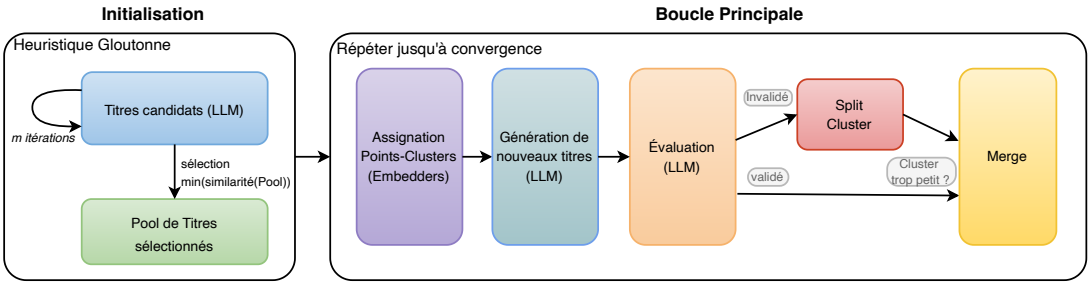


FIGURE 1 – Architecture de JK-Means. La phase d'initialisation repose sur une heuristique greedy de sélection des titres. La boucle principale a une structure séquentielle reprenant les étapes de K-LLMeans en ajoutant un mécanisme d'évaluation par LLM et des opérations de split et merge.

$O(|C_k|)$) tout en identifiant efficacement les clusters hétérogènes. Si un document est invalidé, cela signale que le cluster contient des sous-catégories distinctes nécessitant une subdivision. Cette évaluation est également réutilisée pour définir le critère d'arrêt de l'algorithme (voir Section 3.8).

3.6.2 Mécanisme de Split

Si un cluster C_k a un score de cohérence faible (score$(C_k) < \tau_{split}$) et une taille suffisante ($|C_k| \geq \text{size}_{min}$), nous appliquons une procédure de split pour le diviser en deux sous-clusters. Nous sélectionnons les N_{top} documents les plus proches du centroïde et les N_{bottom} documents les plus éloignés. Pour chaque groupe, le LLM génère un nouveau titre spécifique :

$$t_{k,1} = f_{llm}(\text{“Titre pour : ”} + D^k_{top}) \quad (8)$$

$$t_{k,2} = f_{llm}(\text{“Titre pour : ”} + D^k_{bottom}) \quad (9)$$

Les nouveaux centroïdes sont $\mu_{k,1} = f_{emb}(t_{k,1})$ et $\mu_{k,2} = f_{emb}(t_{k,2})$. Tous les documents de C_k sont ensuite réassignés aux deux nouveaux clusters selon leur affinité. Ce mécanisme permet d'explorer de nouvelles configurations de clusters et de découvrir des sous-catégories sémantiques, palliant ainsi les minima locaux causés par des clusters trop génériques.

3.6.3 Mécanisme de Merge

Un mécanisme de merge est instancié pour éviter les optimums locaux, notamment les clusters de taille 1. En effet, si un cluster ne contient qu'un seul document, le titre généré sera parfaitement adapté à cette unique donnée. Ce cluster ne bougera jamais lors des itérations suivantes, créant ainsi un optimum local qui empêche toute évolution.

De plus, le mécanisme d'exploration engendré par le mécanisme de subdivision évoqué dans la section précédente peut entraîner la création de nombreux clusters de petites tailles. Il faut donc une procédure inverse pour compenser cette fragmentation excessive.

De ce fait, nous supprimons automatiquement les clusters d'une taille inférieur à un seuil (Ce seuil étant un paramètre de l'algorithme voir 4.2). Les documents de ces clusters sont réassignés aux clusters restants selon leur affinité, permettant ainsi de maintenir un équilibre entre exploration (via le split) et consolidation (via le merge).

3.7 Optimisation des Appels LLM

Pour réduire les coûts computationnels liés à l'usage de LLM, nous nous sommes concentré sur quelques leviers d'optimisations :

Assignation au cluster sur une sélection de chunks. Plutôt que de traiter chaque document dans son intégralité pour l'assignation à un cluster, nous calculons le score d'affinité basé sur les $M_{\text{top|chunks}}$ chunks les plus proches d'un centroïde en réalisant la somme des similarités cosinus par cluster et en affectant un document au cluster ayant la plus grande somme sur ces chunks représentatifs. Cette approche capture le cœur sémantique du document sans nécessiter le traitement de l'intégralité du texte, même si en pratique, le paramètre $M_{\text{top|chunks}}$ est réglé de sorte que seuls les documents les plus volumineux d'un dataset soient concernés par cette approximation.

Reconstruction ciblée. Lors de la génération de titres de clusters, nous sélectionnons uniquement les $M_{\text{top|docs}}$ chunks ayant la plus haute similarité cosinus avec le centroïde. Seul le contenu le plus représentatif est envoyé au LLM, réduisant drastiquement le nombre de tokens d'entrée. De façon similaire, ce paramètre intervient uniquement dans le traitement de documents volumineux.

Ces optimisations combinées réduisent significativement les coûts LLM tout en préservant la qualité du clustering.

3.8 Critère d'Arrêt

K-Means ou K-LLMeans sont des approches qui s'arrêtent lorsque le nombre de modifications entre deux itérations est faible (stabilité des assignations), ce qui n'est pas adaptée dans notre cas. En effet, nos mécanismes de split et merge augmentent artificiellement le nombre de modifications effectuées à chaque itération, rendant ce critère d'arrêt inefficace.

Ainsi, nous proposons une méthode plus adaptée qui consiste à réutiliser l'évaluation faite par le LLM as a Judge comme un indicateur de la performance actuelle du modèle.

Nous utilisons cette métrique comme critère d'arrêt en calculant le score de cohérence moyen global :

$$\text{score}_{global} = \frac{1}{K} \sum_{k=1}^{K} \text{score}(C_k) \tag{10}$$

L'algorithme a une patience de 3 itérations (auquel cas nous conservons l'itération avec le meilleur score) ou peut également s'arrêter si aucun split n'est effectué. Dans nos expériences, nous avons utilisé un nombre d'itérations fixes ($N_{\text{iterations}} = 5$) pour garantir la reproductibilité et permettre une comparaison équitable avec les approches antérieures.

4 Expérimentations

4.1 Datasets et Configuration

Nous évaluons notre approche sur trois datasets de référence : 20 Newsgroups [7] (10 classes), AG News [20] (4 classes), et Reuters-21578 [8] (79 classes). Nous utilisons Gemini 2.5 Flash comme LLM et text-embedding-3-large (3072 dimensions) comme modèle d'embedding.

Intentions utilisateur. Pour chaque dataset, nous définissons une intention spécifique guidant le clustering : **20 Newsgroups** ("Classification de groupes de discussion"), **AG News** ("Classification d'articles de presse"), et **Reuters-21578** ("Classification d'articles"). Ces intentions sont intégrées dans tous les prompts LLM (génération de titres initiaux, raffinement, évaluation) pour guider la formation des clusters.

Variantes comparées. Nous comparons deux approches. (1) **K-LLMeans** : K-LLMeans avec prompts intentionnels et sélection des $M_{\text{top|docs}}$ documents les plus proches du centroïde (Section 3.2) pour la génération de titres et initialisé avec K titres avec l'aide d'un LLM. (2) **JK-Means** : Notre approche complète ajoutant l'initialisation gloutonne (Section 3.4), l'évaluation LLM as a Judge (Section 3.6), et les mécanismes de split/merge.

Pour les deux approches, le nombre initial de titres est fixé au nombre de classes du dataset. Cela permet à la première approche d'obtenir les meilleures performances. D'un autre côté, une création trop intensive de nouveaux clusters risque d'impacter négativement la seconde approche en augmentant l'écart avec le nombre de classes réelles.

4.2 Paramètres Expérimentaux

Le Tableau 1 présente les hyperparamètres utilisés pour toutes les expériences. Ces paramètres sont identiques pour les trois datasets afin de garantir une comparaison équitable.

Assignation granulaire. Le paramètre $M_{\text{top|chunks}} = 5$ limite le nombre de chunks les plus pertinents utilisés pour calculer le score d'affinité de chaque document.

Raffinement des titres. Le paramètre $M_{\text{top|docs}} = 5$ limite le nombre de documents envoyés au LLM pour raffiner les titres de clusters. (Dans **K-LLMeans** et **JK-Means**)

Initialisation intentionnelle. Le paramètre $N_{\text{candidates}} = 50$ génère 50 titres candidats via le LLM, parmi lesquels l'heuristique gloutonne sélectionne les K titres les plus diversifiés (où K est le nombre de classes du dataset).

TABLE 1 – Hyperparamètres Expérimentaux

Paramètre	Valeur	Description
$M_{\text{top\|chunks}}$	5	Nombre de chunks les plus pertinents pour l'assignation
$M_{\text{top\|docs}}$	5	Nombre de documents représentatifs pour la génération de titres
$N_{\text{iterations}}$	5	Nombre maximum d'itérations
$N_{\text{max\|chars}}$	10000	Limite de caractères pour les textes envoyés au LLM
$N_{\text{candidates}}$	50	Nombre de titres candidats générés lors de l'initialisation
N_{refine}	1	Fréquence de raffinement des titres via LLM (à chaque itération)
size_{min}	4	Taille minimale d'un cluster avant fusion (merge)

Robustesse aux hyperparamètres. L'utilisation d'un ensemble unique de paramètres pour les trois datasets démontre la robustesse de notre approche. Cette stabilité met en avant la résilience de notre approche face à la perturbation d'un paramètre isolé.

4.3 Métriques d'Évaluation

Nous évaluons notre approche selon cinq métriques complémentaires, permettant l'évaluation de la qualité du clustering ainsi que de la qualité sémantique des titres générés. Nous utilisons des métriques robustes pour l'évaluation de solutions avec un nombre de clusters différents. C'est pourquoi nous prenons une accuracy bijective, ainsi que la NMI et le F1-score.

1. Accuracy Bijective. Cette métrique mesure l'alignement bijectif optimal entre clusters et classes via l'algorithme hongrois [6]. Contrairement à l'accuracy standard, elle pénalise fortement les écarts entre le nombre de clusters générés et le nombre de classes réelles. Ce choix est délibéré : nous souhaitons pénaliser un mécanisme de génération de clusters excessif, s'ils n'apportent que peu d'intérêt.

2. NMI (Normalized Mutual Information). Le NMI [16] mesure l'information mutuelle normalisée entre les clusters prédits et les vraies classes. Cette métrique est robuste aux déséquilibres de classes et quantifie combien d'information sur les vraies classes peut être obtenue en connaissant les clusters.

3. F1-Score. Le F1-Score est la moyenne harmonique de la précision et du rappel, calculée sur les paires de documents. Il équilibre la capacité à créer des clusters homogènes (précision) et à regrouper tous les documents d'une même classe (rappel).

4. Score de Cohérence LLM (Critère d'Arrêt). Comme évoqué en Section 3.8, nous utilisons le score de cohérence moyen calculé par le LLM as a Judge comme indicateur de la qualité du clustering. Ce score mesure le pourcentage de documents validés par le LLM parmi les documents les plus éloignés de chaque centroïde.

5. Similarité Cosinus Maximale (Alignement Sémantique). Pour chaque classe réelle l_j, nous calculons la similarité cosinus maximale avec les titres de clusters générés.

Cette métrique vérifie si nos titres générés sont sémantiquement proches des vraies classes du dataset. Elle illustre la capacité de l'algorithme à découvrir automatiquement des catégories cohérentes avec les labels réels, sans jamais les avoir vus.

5 Résultats et Discussions

5.1 Vue d'Ensemble des Résultats

La Figure 2 présente l'évolution de l'accuracy bijective au fil des itérations pour les trois datasets (AG News, 20 Newsgroups, Reuters). Cette vue globale révèle une tendance claire : JK-Means (courbe bleue) surpasse systématiquement K-LLMeans (courbe orange) sur les trois jeux de données. L'amélioration est particulièrement marquée après quelques itérations, démontrant l'efficacité de la couche exploratoire. Des résultats exhaustifs incluant toutes les métriques pour chaque itération sont présentés en Annexe A (Tableau 5).

Le reste de cette section détaille ces résultats en trois volets complémentaires : (1) une analyse quantitative des performances finales via des métriques de clustering, (2) une évaluation de la qualité de l'initialisation gloutonne, et (3) une mesure de l'alignement sémantique entre les titres générés et les vraies classes. Nous discuterons ensuite les comportements spécifiques observés sur chaque dataset.

5.2 Performances Globales de Clustering

Le Tableau 2 présente les meilleures performances obtenues sur 5 itérations pour les deux approches sur les trois datasets. Les métriques retenues sont l'accuracy bijective, le F1-Score et le NMI.

TABLE 2 – Comparaison des Performances Maximales (Meilleures sur 5 Itérations)

Dataset	Algorithme	Acc. Bij.	F1-Score	NMI
AG News	K-LLMeans	0.6260	0.6430	0.6354
AG News	JK-Means	**0.9600**	**0.9357**	**0.9102**
20NG	K-LLMeans	0.6450	0.5654	0.7563
20NG	JK-Means	**0.7830**	**0.7801**	**0.8821**
Reuters	K-LLMeans	0.6510	0.8252	0.7745
Reuters	JK-Means	**0.7527**	**0.9068**	**0.8258**

Les résultats démontrent des gains significatifs pour l'approche exploratoire sur tous les datasets. Sur AG News, l'amélioration est significative : +53.4% d'accuracy bijective, +45.5% de F1-Score, et +43.3% de NMI. Sur 20 Newsgroups, les gains sont de +21.4% en accuracy, +38.0% en F1-Score, et +16.6% en NMI. Sur Reuters, les améliorations sont de +15.6% en accuracy, +9.9% en F1-Score, et +6.6% en NMI.

Ces résultats montrent l'efficacité de notre approche exploratoire permettant la découverte d'un plus grand nombre de clusters candidats.

5.3 Qualité de l'Initialisation Gloutonne

L'heuristique gloutonne de sélection des titres initiaux vise à maximiser la diversité sémantique et à produire une distribution équilibrée des documents entre les clusters dès la première itération. Le Tableau 3 compare les métriques de distribution initiale entre l'approche standard et l'approche exploratoire.

TABLE 3 – Qualité de la Distribution Initiale (Itération 1)

Dataset	Algorithme	Acc. Bij.	Équilibre
AG News	K-LLMeans	0.4080	0.6961
AG News	JK-Means	**0.5840**	**0.8563**
20 Newsgroups	K-LLMeans	0.2350	0.7413
20 Newsgroups	JK-Means	**0.2840**	**0.8285**
Reuters	K-LLMeans	**0.1343**	**0.7031**
Reuters	JK-Means	0.1103	0.6231

Équilibre : Entropie normalisée [0,1]. Plus élevé = meilleure distribution.

Sur AG News et 20 Newsgroups, l'heuristique gloutonne améliore la qualité de l'initialisation avec des gains respectifs en accuracy de +43.1% et +20.9%. Concernant la mesure d'équilibre, des gains significatifs sur ces mêmes datasets sont observés avec une amélioration de 0.696 à 0.856 sur AG News et de 0.741 à 0.829 pour 20 Newsgroup. Cette stabilité permet de reduire la probabilité d'obtenir un super-cluster (optimum local) lors de l'initialisation.

Le cas de Reuters est plus nuancé : l'accuracy initiale est légèrement inférieure (11.0% vs 13.4%), il en est de même l'équilibre de distribution. Cette apparente contradiction s'explique par la nature du dataset : avec 79 classes, la probabilité d'obtenir un super-cluster générique est très faible, et l'approche standard bénéficie d'une initialisation favorable. Nous reviendrons sur ce point dans la discussion sur Reuters (Section 5.7).

5.4 Alignement Sémantique avec les Vraies Classes

Une métrique cruciale pour évaluer la qualité de notre approche est la similarité cosinus maximale entre les titres de clusters générés et les noms des vraies classes. Cette métrique démontre que l'intention utilisateur guide efficacement la formation des clusters vers des catégories alignées avec les concepts métier, tout en préservant le caractère non supervisé de l'approche. L'évolution détaillée de cette métrique pour chaque itération est présentée dans le Tableau 5 en Annexe A.

JK-Means génère systématiquement des titres plus proches sémantiquement des vraies classes. Sur 20 Newsgroups, à l'itération 4, la similarité cosinus atteint 0.412 (contre 0.324 pour K-LLMeans), soit une amélioration de +27.2%. Les titres générés capturent bien les thématiques réelles (sports, politique, technologie) sans avoir jamais vu les labels. Sur AG News, l'amélioration est encore plus spectaculaire : 0.376 à l'itération 2 (contre 0.273), soit +37.7%, confirmant la capacité de l'algorithme à découvrir des catégories cohérentes (World, Sports, Business, Sci/Tech). Sur Reuters et ces 79 classes, la similarité atteint 0.269 à l'itération 5 (contre 0.202 pour K-LLMeans), soit +33.2%, prouvant que notre approche génère des titres plus représentatifs.

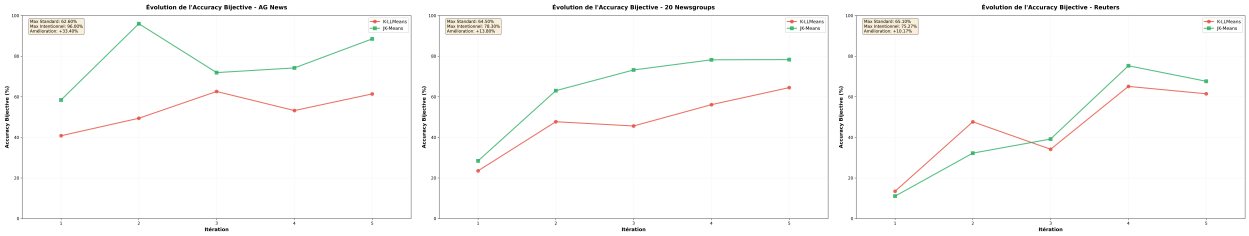


FIGURE 2 – Évolution de l'accuracy bijective au fil des itérations pour les trois datasets (AG News, 20 Newsgroups, Reuters). La courbe orange représente K-LLMeans, tandis que la courbe bleue représente JK-Means. La seconde approche converge rapidement vers des accuracy significativement supérieures, démontrant l'efficacité de l'initialisation et des mécanismes de split/merge.

5.5 Discussion : 20 Newsgroups – Cas Classique de Convergence

Le dataset 20 Newsgroups représente un cas classique où l'algorithme fonctionne de manière optimale. L'algorithme JK-Means surpasse systématiquement l'autre approche avec des gains significatifs : +21.4% en accuracy bijective (78.3% vs 64.5%), +38.0% en F1-Score (0.780 vs 0.565), et +16.6% en NMI (0.882 vs 0.756). Cette convergence rapide et stable valide l'efficacité de notre approche sur des données avec une structure sémantique claire et un nombre modéré de classes.

5.6 Discussion : AG News – Sur-Fragmentation et Critère d'Arrêt

Le dataset AG News présente un comportement particulier qui révèle une limitation de notre brique exploratoire : la tendance à la sur-fragmentation sur des datasets avec un nombre réduit de classes. AG News ne contient que 4 classes (World, Sports, Business, Sci/Tech), ce qui rend l'algorithme vulnérable à la création excessive de clusters via le mécanisme de split.

5.6.1 Évolution du Nombre de Clusters

Le Tableau 4 montre l'évolution du nombre de clusters de notre approche au fil des itérations. L'algorithme démarre avec 4 clusters, puis augmente progressivement jusqu'à 13 clusters à l'itération 4, avant de redescendre à 10 à l'itération 5.

TABLE 4 – Évolution du Nombre de Clusters sur AG News, pour la variante exploratoire

Itération	Nombre de Clusters	Acc. Bijective
1	4	0.5840
2	5	**0.9600**
3	9	0.7190
4	13	0.7420
5	10	0.8850

Cette sur-fragmentation explique la chute d'accuracy observée à partir de l'itération 3. L'accuracy bijective pénalise les écarts entre le nombre de clusters générés et le nombre de classes réelles. À l'itération 2, avec 5 clusters, l'accuracy atteint un pic optimal de 96.0%. Lors de l'itération suivante, l'augmentation du nombre de clusters entraine une chute de l'accuracy (71.9%).

5.6.2 Utilisation du Score de Cohérence comme Critère d'Arrêt

Le Tableau 5 en annexe présente l'évolution du score de cohérence LLM as a Judge en parallèle des métriques de performance. En s'arrêtant à l'itération où le score de cohérence est maximal, on éviterait la chute d'accuracy des itérations suivantes. Cette observation se vérifie sur les trois datasets : sur 20 Newsgroups et Reuters, l'itération avec le score de cohérence maximal correspond effectivement à la meilleure accuracy obtenue, tandis que sur AG News, elle correspond au second meilleur score, très proche de l'optimum. Ce critère d'arrêt met en avant un second intérêt de notre évaluation LLM as a Judge.

5.7 Discussion : Reuters – Initialisation Défavorable et Rattrapage

Le dataset Reuters présente un comportement inverse à celui d'AG News : l'approche K-LLMeans surperforme initialement, mais JK-Means rattrape et dépasse sur le long terme. Une explication se base sur le haut nombre de classes (79) dans ce dataset.

Avec une initialisation à 79 clusters, l'approche K-LLMeans a peu de chance d'avoir une initialisation défavorable. La chute de performance à l'itération suivante peut même présumer d'une initialisation favorable. D'un autre côté, notre heuristique gloutonne présente des faiblesses dans ce cas de figure. Sur une base de 250 titres générés (uniquement sur ce jeu de données) via un LLM pour n'en sélectionner que 79, il est possible qu'une aussi haute génération de titres basée sur une intention puisse générer de nombreux titres peu informatifs. De ce fait, notre heuristique gloutonne pourrait prioriser des titres avec une richesse sémantique trop faible.

Cependant, dès l'itération 2, JK-Means rattrape (32.2% vs 47.7%), puis dépasse progressivement K-LLMeans. À l'itération 4, nous atteignons 75.3% d'accuracy, contre 65.1%. Cette évolution démontre que les mécanismes d'exploration (split/merge) de notre approche permettent de compenser un mauvais point de départ, prouvant la robustesse de l'algorithme face aux conditions initiales défavorables.

6 Conclusion et Perspectives

Nous avons proposé JK-Means, une approche de clustering non supervisé guidé par l'intention utilisateur exprimée en langage naturel. Notre méthode repose sur trois contributions principales : (1) une initialisation gloutonne maximisant la diversité sémantique des titres initiaux, (2) une évaluation LLM as a Judge en temps constant favorisant l'exploration de nouveaux clusters, et (3) des mécanismes de split/merge permettant d'adapter dynamiquement le nombre de clusters.

Les expérimentations sur trois datasets de référence démontrent des gains significatifs en accuracy bijective, F1-Score et NMI sur tous les datasets. Plus important, notre métrique de similarité sémantique prouve que l'algorithme découvre automatiquement des catégories cohérentes avec les vraies classes, sans jamais les avoir vues. Cette capacité ouvre la voie à des applications métier où l'expertise utilisateur peut guider la découverte de structures dans les données.

Perspectives. Nos expérimentations révèlent deux écueils liés au nombre de classes. Sur des datasets avec un nombre très faible de classes (AG News, 4 classes), le mécanisme de split tend à créer une sur-fragmentation. Bien que notre critère d'arrêt basé sur le score de cohérence LLM permette d'éviter ces problèmes en détectant le point optimal avant la dégradation, une régulation plus fine des mécanismes de split/merge pourrait améliorer la stabilité. À l'inverse, sur des datasets avec un très grand nombre de classes (Reuters, 79 classes), la probabilité d'obtenir un super-cluster générique lors de l'initialisation devient très faible, rendant notre heuristique gloutonne moins déterminante.

Une perspective actuellement en cours est d'intégrer de l'exploitation dans la phase de split. Plutôt que de simplement diviser un cluster hétérogène en deux sous-clusters basés sur les documents les plus proches et les plus éloignés du centroïde, nous pourrions réeffectuer un algorithme de type K-LLMeans à petite échelle entre toutes les données les plus éloignées des clusters entre eux. Cette approche favoriserait un rapprochement de données similaires et permettrait d'optimiser le mécanisme de subdivision en réduisant la génération de nouveaux clusters de taille réduite.

Remerciements

Les auteurs remercient chaleureusement Talan pour son soutien constant et son engagement dans ces travaux de recherche. L'ensemble du code source est disponible en open source à l'adresse suivante : https://github.com/ArnaudDeleruyelle/JKMeans.

A Résultats Détaillés par Itération

Le Tableau 5 présente l'évolution complète de toutes les métriques pour les deux approches sur les trois datasets au fil des 5 itérations.

B Prompts utilisés

Cette annexe présente les trois prompts principaux utilisés dans notre approche K-LLMeans et JK-Means. Les variables sont indiquées entre crochets (ex : `[INTENTION]`, `[NUM_CANDIDATES]`).

B.1 Prompt 1 : Génération des Titres Initiaux

Prompt de Génération des Titres Initiaux

Tu es un expert en analyse sémantique de documents. Ta tâche est de proposer une liste très diverse de `[NUM_CANDIDATES]` sous-catégories spécifiques liées au thème principal : '`[INTENTION]`'.
Suis ces règles strictes pour chaque titre :

- Sois très spécifique et distinct des autres titres.
- Doit être concis (3 à 6 mots).
- **N'utilise pas de termes génériques** comme 'Système', 'Erreur', 'Problème', ou 'Incident'. Concentre-toi sur des sujets techniques ou fonctionnels concrets.
- Varie au maximum les domaines fonctionnels et techniques.

Retourne uniquement la liste des titres, un par ligne, sans numéros ni puces.

B.2 Prompt 2 : Régénération des Titres de Clusters

Prompt de Régénération des Titres

Tu es un expert en analyse sémantique de documents. Analyse les documents suivants pour en extraire le thème commun et créer un titre de cluster.
— Règles Impératives —

1. **Sois Unique** : Le nouveau titre doit être sémantiquement distinct des titres existants listés ci-dessous.
2. **Sois Spécifique** : Évite les termes trop génériques comme 'Problème', 'Erreur', 'Système'.
3. **Sois Concis** : Le titre doit faire entre 3 et 6 mots.
4. **Respecte le Thème** : Le titre DOIT s'aligner avec le thème général : '`[INTENTION]`'.

— Documents à Analyser —
Document 1 (Titre Original : '`[DOC_TITLE_1]`') : — Début du Document — `[CONTENU_1]` — Fin du Document —
Document 2 (Titre Original : '`[DOC_TITLE_2]`') : — Début du Document — `[CONTENU_2]` — Fin du Document —
...
— Titres Existants (à ne pas imiter) — - `[TITRE_EXISTANT_1]` - `[TITRE_EXISTANT_2]` - ...
IMPORTANT : Ne retourne STRICTEMENT que le nouveau titre commun et unique, et rien d'autre.
Nouveau Titre :

TABLE 5 – Évolution complète des métriques (sans ARI) par itération

Dataset	Algorithme	It.	Acc.	F1	NMI	Coh.	Cos.	Dist.
AG News	K-LLMeans	1	0.408	0.366	0.143	-	0.115	1.330
AG News	K-LLMeans	2	0.494	0.467	0.375	-	0.273	1.203
AG News	K-LLMeans	3	0.626	0.584	0.487	-	0.229	1.241
AG News	K-LLMeans	4	0.532	0.558	0.493	-	0.242	1.230
AG News	K-LLMeans	5	0.614	0.643	0.635	-	0.269	1.208
AG News	JK-Means	1	0.584	0.480	0.286	50.7%	0.115	1.331
AG News	JK-Means	2	**0.960**	**0.936**	**0.910**	**69.8%**	**0.376**	**1.116**
AG News	JK-Means	3	0.719	0.743	0.735	43.7%	0.361	1.128
AG News	JK-Means	4	0.742	0.818	0.812	55.3%	0.344	1.145
AG News	JK-Means	5	0.885	0.844	0.793	59.5%	0.333	1.154
20 News	K-LLMeans	1	0.235	0.279	0.308	-	0.100	1.341
20 News	K-LLMeans	2	0.477	0.471	0.632	-	0.258	1.210
20 News	K-LLMeans	3	0.456	0.414	0.657	-	0.286	1.185
20 News	K-LLMeans	4	0.561	0.533	0.743	-	0.324	1.153
20 News	K-LLMeans	5	0.645	0.565	0.756	-	0.294	1.180
20 News	JK-Means	1	0.284	0.307	0.371	18.4%	0.113	1.332
20 News	JK-Means	2	0.630	0.573	0.730	43.0%	0.365	1.122
20 News	JK-Means	3	0.732	0.752	0.863	65.2%	0.392	1.098
20 News	JK-Means	4	0.782	0.721	0.862	**85.2%**	**0.412**	**1.079**
20 News	JK-Means	5	**0.783**	**0.780**	**0.882**	76.8%	0.393	1.096
Reuters	K-LLMeans	1	0.134	0.336	0.442	-	0.128	1.320
Reuters	K-LLMeans	2	0.477	0.640	0.712	-	0.185	1.273
Reuters	K-LLMeans	3	0.341	0.619	0.700	-	0.191	1.269
Reuters	K-LLMeans	4	0.651	0.825	0.775	-	0.196	1.264
Reuters	K-LLMeans	5	0.615	0.753	0.766	-	0.202	1.258
Reuters	JK-Means	1	0.110	0.561	0.536	43.6%	0.160	1.296
Reuters	JK-Means	2	0.322	0.558	0.659	57.2%	0.253	1.219
Reuters	JK-Means	3	0.392	0.637	0.726	53.7%	**0.279**	**1.198**
Reuters	JK-Means	4	**0.753**	**0.907**	**0.826**	**70.2%**	0.263	1.210
Reuters	JK-Means	5	0.676	0.816	0.796	66.4%	0.269	1.206

It. : Itération. **Acc.** : Accuracy bijective. **F1** : F1-Score. **NMI** : Normalized Mutual Information. **Coh.** : Score de cohérence LLM (critère d'arrêt). **Cos.** : Similarité cosinus max. **Dist.** : Distance euclidienne min.

B.3 Prompt 3 : Évaluation LLM as a Judge

Prompt d'Évaluation LLM as a Judge

L'objectif est de regrouper les documents selon le thème suivant : `[INTENTION]`.
Contenu du document : `[DOCUMENT]`
Titre du cluster : `[TITRE_CLUSTER]`
Le titre du cluster correspond à une catégorie générale ou fonctionnelle. Votre objectif est d'évaluer si ce titre est **pertinent** pour regrouper ce document, même s'il ne couvre pas tous les détails du résumé, en tenant compte du thème général défini ci-dessus.
Considérez le titre comme **représentatif** si :

- Il reflète le thème principal du document (même partiellement).
- Il désigne un module, un processus, ou une activité en lien clair avec le résumé.
- Il est suffisamment proche pour qu'un regroupement de documents proches soit possible.
- Il respecte l'intention de clustering définie.

Répondez uniquement par 'OUI' si le lien est **raisonnable et justifié**, même s'il n'est pas strict ou exhaustif. Sinon, répondez 'NON'.

Références

[1] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Proceedings of the 8th International Conference on Database Theory (ICDT)*, volume 1973 of *Lecture Notes in Computer Science*, pages 420–434. Springer, 2001.

[2] David Arthur and Sergei Vassilvitskii. k-means++ : The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035, 2007.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.

[4] Jairo Diaz-Rodriguez. Summaries as centroids for interpretable and scalable text clustering. *arXiv preprint arXiv :2502.09667*, 2025.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, 1996.

[6] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2) :83–97, 1955.

[7] Ken Lang. Newsweeder : Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, pages 331–339, 1995.

[8] David D. Lewis. Reuters-21578 text categorization test collection, distribution 1.0. http://www.daviddlewis.com/resources/testcollections/reuters21578/, 1997.

[9] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2) :129–137, 1982.

[10] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[12] Justin K. Miller and Tristram J. Alexander. Human-interpretable clustering of short text using large language models. *Royal Society Open Science*, 12(1), 2025.

[13] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove : Global vectors for word representation. In *EMNLP*, 2014.

[14] Chau Minh Pham, Alexander Hoyle, Simran Srinivasan, and Nils Reimers. Topicgpt : A prompt-based topic modeling framework. In *NAACL*, 2024.

[15] Nils Reimers and Iryna Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[16] Andrew Rosenberg and Julia Hirschberg. V-measure : A conditional entropy-based external cluster evaluation measure. pages 410–420, 2007.

[17] Vijay Viswanathan, Kiril Tomkins, Arjun Mahajan, Ellie Pavlick, and Chris Callison-Burch. Large language models enable few-shot clustering. *arXiv preprint arXiv :2307.00524*, 2023.

[18] Hongtao Wang, Renchi Yang, and Wenqing Lin. Nilc : Discovering new intents with llm-assisted clustering. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pages 671–680, 2026.

[19] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm : Deep self-attention distillation for task-agnostic compression. In *NeurIPS*, 2020.

[20] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 649–657, 2015.

[21] Yuwei Zhang, Zihan Meng, Preslav Nakov, and Sophia Ananiadou. Clusterllm : Large language models as a guide for text clustering. *arXiv preprint arXiv :2305.14871*, 2023.

[22] Hui Zheng, Yongfeng Liu, Xuanli Yang, Kexun Chen, and Yu-Ping Tan. Systematic evaluation of llm-as-a-judge in llm alignment tasks. *arXiv :2408.13006*, 2024.

Evaluer la créativité analogique - Une expérimentation

Stergos Afantenos[1], Leonardo Cortez Bernardes[1], Henri Prade[1]

[1] Institut de Recherche en Informatique de Toulouse (IRIT), Univ. Toulouse, Toulouse INP, CNRS

stergos.afantenos@irit.fr, bernardesleonardo0@gmail.com, henri.prade@irit.fr

Résumé

Le raisonnement analogique est considéré comme créatif. Après avoir examiné les différents points de vue sur la créativité computationnelle et son évaluation, nous nous intéressons à une forme particulière d'inférence analogique basée sur les proportions analogiques, qui sont des relations quaternaires comparant les différences entre deux paires d'éléments. Ce mécanisme d'inférence est capable de générer un nouvel élément à partir de trois éléments décrits par le même ensemble de caractéristiques. Ce mécanisme est utilisé, dans le cadre d'une expérimentation, pour générer de nouveaux prompts adressés à un chatbot basé sur un LLM afin de générer du texte contraint. Les prompts spécifient les émotions à l'œuvre dans un court texte décrivant une situation banale. La distance de Hellinger et la divergence de Kullback-Leibler sont utilisées pour des évaluations de la créativité au niveau du prompt ou du texte. Les mérites relatifs de ces deux mesures d'évaluation de la créativité sont discutés. L'article s'appuie sur un mécanisme créatif original couplant l'inférence analogique et la génération de texte basée sur le LLM.

Mots-clés

Créativité, proportion analogique, LLM.

Abstract

Analogical reasoning is said to be creative. After discussing views of computational creativity and its evaluation, we consider a particular form of analogical inference based on analogical proportions, which are quaternary relations comparing the differences inside two pairs of items. This inference mechanism is capable of generating a new item from three items described on the same set of features. This mechanism is used, in an experiment, to generate new prompts that are addressed to an LLM-based chatbot to generate constrained text. The prompts specify the emotions at work in a short text describing a mundane situation. Hellinger distance and Kullback-Leibler divergence are used for provisional assessments of creativity at the prompt level or the text level. The relative merits of these two measures of creativity assessment are discussed. The paper relies on an original creative mechanism coupling analogical inference and LLM-based text generation.

Keywords

Creativity, analogical proportion, LLM.

1 Introduction

La créativité désigne la capacité à développer des idées ou des productions originales et, de préférence, de valeur. Cela s'applique clairement à un large éventail d'œuvres. La créativité humaine englobe les arts, la littérature et l'humour, ainsi que la résolution de problèmes, laquelle couvre une variété de situations rencontrées en mathématiques et en sciences, ainsi que dans les problèmes de la vie quotidienne. La créativité, en tant qu'atout majeur de l'esprit humain, constitue depuis longtemps un enjeu important de la recherche en intelligence artificielle [8, 41, 18], comme on peut le voir notamment pour la musique [69], les arts visuels [19] et la littérature [64].

Le recours aux analogies est reconnu comme une forme de créativité, en particulier pour la résolution de problèmes. Cette approche a été préconisée par de nombreux auteurs, en psychologie cognitive (par exemple, [22, 23, 32]), en intelligence artificielle [65] et en mathématiques [46]. La créativité apparaît d'autant plus évidente et frappante que l'analogie rapproche deux situations éloignées, ou deux cadres de référence tout à fait différents.

Le raisonnement analogique peut prendre diverses formes selon les cadres de modélisation utilisés. En "Structure-mapping theory" et dans les approches connexes [30, 26, 34], le raisonnement analogique est considéré comme la capacité à établir des parallèles entre deux univers (ou, plus simplement, deux situations) en faisant correspondre les relations entre les éléments du premier univers (et leurs propriétés) à celles applicables aux éléments du second univers, et à tirer des conclusions potentielles dans le second univers à partir de la similarité des relations dans les deux univers (le second univers étant supposé ici moins connu que le premier). Ce type de raisonnement analogique peut être encodé dans un cadre logique classique (nécessitant une anti-unification du second ordre) [31] où l'on peut, par exemple, saisir la célèbre analogie créative de Rutherford, où on compare le système solaire (jouant le rôle du premier univers) au modèle atomique de Rutherford (le second univers) [62].

Le raisonnement à partir de cas [38, 1, 28] est souvent considéré comme une forme simple de raisonnement analogique. Il utilise un répertoire de cas, composés de paires (problème, solution), et propose une solution à un problème cible en adaptant la solution d'un problème similaire présent dans ce répertoire. Cependant, la créativité du raison-

nement à partir de cas demeure relativement modeste, dans la mesure où les problèmes comparés sont similaires[1]. Un parallèle est ainsi établi entre la paire (problème, solution) utilisée et le problème cible, ainsi que sa solution proposée. On peut dire que la solution proposée est au problème cible ce que la solution du cas similaire est au problème du cas présenté.

L'énoncé ci-dessus, qui relie quatre éléments, est une proportion analogique de la forme « a est à b ce que c est à d ». En effet, lorsqu'on établit un parallèle entre une paire (a, b) dans un premier univers et la paire correspondante (c, d) dans un second univers, on obtient de telles proportions, comme l'indique [33]. Par exemple, dans l'analogie de Rutherford, « l'électron est au noyau ce que la planète est au soleil ». Dans cet exemple, la proportion analogique implique deux univers (très différents) ; dans le raisonnement à partir de cas, on observe une autre distinction entre problèmes et solutions.

Il existe un troisième type de raisonnement analogique, directement fondé sur les proportions analogiques. Sachant que « a est à b ce que c est à d », on infère d (ou une partie inconnue de sa description) à partir du triplet (a, b, c). Ce raisonnement exige que les quatre éléments soient de même nature, c'est-à-dire qu'ils puissent être décrits par le même ensemble d'attributs, ce qui n'est pas nécessairement le cas dans les deux approches précédentes. Il convient de souligner [4] que les proportions analogiques peuvent soit impliquer quatre éléments descriptibles par le même ensemble d'attributs, comme dans « un chiot est à un chien ce qu'un chaton est à un chat », soit associer deux univers (celui de a et c, et celui de b et d) qui se réfèrent à deux ensembles d'attributs distincts, comme dans « le vin est aux Français ce que la bière est aux Anglais ». Dans cet article, nous nous concentrons uniquement sur le premier type de proportions analogiques.

Les proportions analogiques semblent créatives dans la mesure où a, b, c et d peuvent être tous différents, et où d – qui peut être obtenu à partir de a, b, c – est décrit par une recombinaison des caractéristiques de a, b, c. Cependant, la question de l'évaluation d'une telle créativité se pose. Dans cet article, nous proposons une expérience qui illustre ce processus créatif dans le domaine de la génération de texte, et nous tentons de mesurer la nouveauté du résultat obtenu.

Cet article[2] étudie la créativité à l'aide de l'inférence basée sur les proportions analogiques. Il est organisé comme suit. La Section 2 passe en revue différents aspects de la créativité et les méthodes pour l'évaluer. Elle examine également les moyens de mesurer la créativité qui ont été proposés, tels que la distance de Hellinger ou la divergence de Kullback-Leibler, et discute de leurs mérites respectifs. La Section 3 présente les proportions analogiques et l'inférence analogique, et souligne leur potentiel pour la créativité, qui est évalué théoriquement. La Section 4 décrit le cadre d'une expérience : l'inférence basée sur les proportions analogiques est utilisée pour générer des combinaisons de caractéristiques relatives à un court texte rapportant une suite de faits ordinaires. Ces combinaisons de caractéristiques servent ensuite d'invites ($prompts$) adressées à un chatbot basé sur un LLM pour générer des textes contraints. Ainsi, la créativité peut être mesurée à la fois au niveau des prompts et au niveau du texte. La Section 5 présente les résultats de l'expérience, et la Section 6 propose une discussion finale.

1. Cependant, à la suite d'idées qui apparaissent déjà dans [59], il existe des travaux sur la créativité basés sur le raisonnement à partir de cas ; citons dans le domaine musical [51, 52], [21].

2. Une version anglaise de cet article a été acceptée à la 17th International Conference on Computational Creativity [2].

2 La créativité et son évaluation

Dans cette section, nous passons d'abord en revue les travaux sur la créativité computationnelle visant à la formaliser ou à l'évaluer. Nous rappelons ensuite deux mesures statistiques de divergence proposées pour évaluer la nouveauté, un aspect majeur de la créativité.

2.1 Revue de travaux en créativité computationnelle

La créativité computationnelle désigne les systèmes d'IA capables d'imiter les processus créatifs humains. Ceci nous ramène à la créativité humaine. C'est pourquoi il peut être utile de partir de la définition de la créativité (humaine) donnée dans un article très cité [53] : "The word creativity is a noun naming the phenomenon in which a person communicates a new concept (which is the product). Mental activity (or mental process) is implicit in the definition, and of course no one could conceive of a person living or operating in a vacuum, so the term *press* is implicit. The definition begs the questions as how new the concept must be and to whom it must be new." Cette première approche est connue pour introduire quatre « p » : personne, produit, processus et presse, impliqués dans tout processus créatif.

La créativité informatique fait l'objet de recherches actives depuis plus de trente ans. De manière générale, outre les travaux portant sur des programmes informatiques spécifiques capables de produire de nouvelles œuvres dans les domaines de l'art, de la littérature ou des mathématiques, de nombreux travaux ont tenté de définir, de catégoriser et de formaliser la créativité informatique en général, ou de proposer des méthodes générales d'évaluation de la créativité.

Cela a débuté avec les travaux influents de Boden [8, 9], qui distinguait trois types de créativité : combinatoire, exploratoire et transformationnelle. Dans la créativité combinatoire, le résultat est obtenu par une simple combinaison d'idées classiques, tandis que les deux autres formes de créativité se déploient dans un espace conceptuel : la première explore cet espace, tandis que la seconde le transforme (par exemple, en assouplissant certaines des contraintes qui le définissent). La créativité transformationnelle est considérée comme la forme la plus aboutie de créativité, dépassant la simple nouveauté.

Bundy [13] souligne que, selon Boden [9], un démonstrateur de théorèmes automatique résolvant une conjecture ouverte en mathématiques ne serait pas considéré comme vé-

ritablement créatif, le nouveau résultat étant obtenu par une procédure exploratoire. En revanche, lorsqu'une conjecture ouverte est résolue par un humain, son travail est vraisemblablement considéré comme créatif. Bundy a insisté sur la nécessité de prendre en compte la complexité potentiellement importante de l'exploration de l'espace conceptuel. Ajoutons qu'en mathématiques, outre la difficulté du résultat à établir, une démonstration novatrice fondée sur une nouvelle approche du problème peut être considérée comme un exemple de créativité transformationnelle.

En outre, Boden [9] fait la distinction entre la créativité p (« p » pour psychologique) et la créativité h (« h » pour historique), où la première est jugée par rapport aux productions de l'agent, tandis que le résultat de la seconde créativité est considéré comme nouveau par la société.

Les travaux inspirants de Boden ont incité certains chercheurs à formaliser et à affiner ses définitions d'espace conceptuel ou de créativité transformationnelle (voir par exemple [55] pour une discussion à ce sujet). Parmi les pistes de recherche notables, citons le développement d'une « théorie formelle de la créativité » par Wiggins [66, 67, 68]), où l'espace conceptuel est appréhendé comme un ensemble d'artefacts possibles, associés à un ensemble de règles qui sélectionnent ceux qui sont acceptables (jouant ainsi le rôle d'un contrôle qualité), et où un autre ensemble de règles attribue une valeur à un artefact créé [67].

Wiggins conçoit la créativité transformationnelle comme une créativité exploratoire à un niveau méta, reprenant une idée défendue dans [12]. La créativité est envisagée comme un processus de recherche généralisé : i) sans hypothèse de monde clos ; ii) permettant la combinaison d'idées existantes ; iii) distinguant les heuristiques qui évaluent la qualité des solutions de celles utilisées pour trouver efficacement une solution ; et iv) privilégiant la notion de comparaison à celle de valeur. Cette conception de la créativité comme recherche a été approfondie et reformulée dans [57]. Plus récemment, le cadre de la théorie formelle de la créativité a été utilisé pour affirmer que « si la nouveauté n'est ni nécessaire ni suffisante à la créativité transformationnelle en général, lorsqu'un agent utilise un ensemble inspirant, plutôt qu'une séquence d'expériences, la nouveauté est effectivement requise pour que la créativité transformationnelle se manifeste » [58].

Une autre piste de recherche, inspirée par les travaux de Boden, est l'approche plus pragmatique proposée par Ritchie [54, 56], qui propose une série de critères pour attribuer de la créativité à un programme informatique, fondée sur trois indicateurs. En effet, outre les deux indicateurs reconnus de créativité que sont la *nouveauté* (l'élément produit doit être différent de ce qui est connu) et la *qualité* (l'élément produit doit avoir une valeur dans son genre), Ritchie ajoute la *typicité* (l'élément produit doit être représentatif de sa classe). De plus, le programme créatif peut utiliser un *ensemble inspirant* qui est conçu pour pouvoir recréer différemment (dans la mesure du possible). Les artefacts de sortie à forte valeur ajoutée d'un programme créatif, qui ne font pas déjà partie de l'ensemble inspirant, constituent l'*ensemble créatif* [17].

Maher [40] soutient que l'évaluation de la créativité est une combinaison de l'évaluation de trois critères essentiels de créativité : la nouveauté, la valeur et l'inattendu (ou la surprise), où « l'inattendu est mesuré par la distance entre l'artefact et l'artefact suivant attendu » (en tenant compte de la nature séquentielle de l'attente).

Schmidhuber [61] adopte une perspective d'apprentissage par renforcement pour modéliser les processus créatifs, où « la récompense créative du sélecteur d'actions correspond à la surprise ou à la nouveauté actuelle, mesurée par les améliorations apportées au modèle du monde ». L'idée est d'utiliser l'apprentissage par renforcement pour « créer, découvrir et apprendre de nouveaux schémas dans l'historique croissant des actions et des entrées sensorielles, la théorie spécifiant formellement ce qu'est un schéma, ce qui est nouveau ou surprenant, et ce que signifie l'apprentissage progressif de nouvelles compétences menant à des schémas encore plus inédits ». In [60] Schmidhuber défend l'idée que la créativité est un sous-produit d'un processus simple de compression de l'information qui privilégie les formes régulières moins complexes à coder (et trouvées plus belles).

Il ne faut pas confondre cela avec l'utilisation d'outils d'apprentissage automatique modernes pour la créativité computationnelle, où, par exemple, des images surprenantes peuvent être générées par un modèle de conversion texte-image à partir de consignes. Dans ce cas, la créativité repose sur la consigne (généralement d'origine humaine) qui spécifie la nécessité d'une *combinaison* inhabituelle d'éléments qui ne se rencontrent pas ensemble dans le monde réel.

Enfin, évoquons quelques idées générales souvent associées à la créativité. La créativité résulte fréquemment d'une fusion conceptuelle [63], qui correspond à l'association de deux cadres de référence différents pouvant être joints, reliés. Le raisonnement analogique, notamment lorsqu'il établit un parallèle entre deux univers distincts par le biais d'un $structure\ mapping$, est souvent considéré comme une forme de fusion conceptuelle ($conceptual\ blending$).

La « pensée latérale » [20], en résolution de problèmes, consiste à passer d'un modèle familier à un modèle nouveau et inattendu. Ceci se rapproche de l'idée de « penser hors des sentiers battus », qui s'apparente à une opération de transformation visant à élargir et à modifier l'espace conceptuel.

2.2 Mesures potentielles de nouveauté

L'étude précédente a montré que tous s'accordent à dire que la nouveauté est un critère essentiel d'évaluation de la créativité. La qualité et la valeur des éléments produits semblent plus faciles à maintenir grâce aux contraintes imposées par le programme informatique. Dans cette section, nous passons en revue deux mesures utilisées en pratique pour évaluer la nouveauté : la distance de Hellinger et la divergence de Kullback-Leibler.

Ces mesures comparent deux distributions de probabilité. Appliquées à la créativité, on peut considérer que la première distribution décrit l'ensemble des éléments inspirants

et la seconde, les éléments produits par le système. Ces mesures évaluent dans quelle mesure les éléments produits diffèrent des exemples initiaux.

2.2.1 Distance de Hellinger

Cette distance est utilisée dans l'évaluation de la créativité, par exemple dans [27]. Elle a été introduite pour la première fois en 1909 par Hellinger, un mathématicien allemand. Étant donné deux distributions de probabilité discrètes $P = (p_1, \ldots, p_n)$ et $Q = (q_1, \ldots, q_n)$ sur le même espace $X = \{x_1, \ldots, x_n\}$ (où $p_i = P(\{x_i\})$, $\sum_{i=1,n} p_i = 1 = \sum_{i=1,n} q_i$), leur distance de Hellinger est définie par

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1,n} (\sqrt{p_i} - \sqrt{q_i})^2}$$

On peut voir facilement que

$$H^2(P,Q) = 1 - \sum_{i=1,n} \sqrt{p_i q_i}$$

Comme on peut le constater, la distance de Hellinger est liée à la distance euclidienne. Nous avons $0 \leq H(P,Q) \leq 1$. $H(P,Q) = 0$ si et seulement si $p_i = q_i$ pour tout i. $H(P,Q) = 1$ si et seulement si P and Q sont mutuellement singulières, i.e., $\forall i, p_i > 0 \Rightarrow q_i = 0$ et inversement $\forall i, q_i > 0 \Rightarrow p_i = 0$.
La quantité $\sum_{i=1,n} \sqrt{p_i q_i}$ est connue comme le coefficient de Bhattacharyya, une mesure bien connue de similarité entre deux distributions de probabilité P et Q. Ainsi, $H(P,Q)$ est une mesure de dissimilarité entre ces distributions.

2.2.2 La divergence de Kullback-Leibler

Cette divergence [39] est une mesure d'entropie relative qui évalue l'écart entre la distribution de probabilité Q et la distribution de probabilité P. Elle est définie par

$$KL(P||Q) = \sum_{i=1,n} p_i \ln \frac{p_i}{q_i} = \sum_{i=1,n} (p_i \ln p_i - p_i \ln q_i)$$

Généralisation naturelle de l'entropie de Shannon, cette mesure a été caractérisée de différentes manières et appliquée dans divers domaines. Elle est positive : $KL(P||Q) \geq 0$, La divergence est non symétrique $KL(P||Q) \neq KL(Q||P)$, Il ne s'agit donc pas d'une distance. L'utilisation classique de cette divergence consiste à évaluer dans quelle mesure Q constitue une approximation acceptable de P. Dans cette perspective, plus la divergence est faible, mieux c'est : plus P et Q sont similaires. $KL(Q||P) = 0$ si et seulement si $p_i = q_i$ for all i. Il ne s'agit donc pas d'une distance. L'utilisation classique de cette divergence est de juger dans quelle mesure Q est une approximation acceptable de P. Dans cette perspective, plus la divergence est faible, mieux c'est, car P et Q sont similaires. $KL(Q||P) = 0$ si et seulement si $p_i = q_i$ pour tout i.

De plus, $KL(P||Q)$ est fini uniquement si le support de P est inclus dans le support de Q ($\forall i, p_i > 0 \Rightarrow q_i > 0$). En revanche, si $\exists i, p_i > 0$ et $q_i = 0$, alors $KL(P||Q)$ est infini. Ainsi, lorsque $KL(P||Q) > 0$ (et n'est pas proche de 0), cela indique que Q est dissemblable par rapport à P.

La divergence de Kullback-Leibler a été proposée comme une mesure de « surprise bayésienne » [36]. Elle a été utilisée pour détecter la surprise dans des récits courts, démontrant qu'une modification de la prédiction du modèle correspond à une modification significative de l'évaluation des participants [15]. De plus, son utilisation a permis d'identifier des auteurs et des œuvres clés qui définissent les grandes orientations de l'évolution littéraire dans un vaste corpus de romans français du XIX^e et du début du XX^e siècle [5]. L'utilisation de cette divergence semble pertinente pour évaluer la nouveauté.
Enfin, certains auteurs ont proposé d'utiliser la complexité algorithmique comme critère d'attribution de la créativité [42, 45, 11]. La complexité de Kolmogorov conditionnelle est proposée pour évaluer la nouveauté. L'idée est qu'un artefact est de haute qualité s'il représente le fruit d'un effort de calcul significatif. Bien que théoriquement séduisante, cette approche est difficile à mettre en œuvre, même si des évaluations approximatives sont possibles en principe. Nous n'avons donc pas retenu cette option pour notre expérience, malgré le fait que des proportions analogiques puissent également être envisagées en termes de complexité de Kolmogorov [7, 47].

3 Inférence analogique et créativité

Nous rappelons d'abord la notion de proportions analogiques et leur modélisation booléenne, avant d'expliquer comment les utiliser à des fins d'inférence et de création. Nous concluons cette section par une étude théorique de leur potentiel créatif.

3.1 Proportions analogiques

Les proportions analogiques sont des énoncés reliant quatre entités de même nature. Ils sont de la forme « a est à b ce que c est à d ». En principe, ces entités peuvent être de toute nature : nombres, mots, images, phrases, etc. Les proportions analogiques sont apparues pour la première fois (du moins en Occident) avec Aristote, qui s'est inspiré des travaux de mathématiciens de son époque sur les proportions arithmétiques ($a - b = c - d$) et géométriques ($\frac{a}{b} = \frac{c}{d}$). Ces proportions numériques égalisent respectivement les différences et les rapports. Aristote a étendu ce principe aux mots (ou concepts).
Comme on peut le constater, de telles proportions opèrent à la fois une comparaison au sein de paires et une comparaison entre paires. C'est précisément ce que signifie l'énoncé « a est à b ce que c est à d ». Une telle proportion est notée $a : b :: c : d$. Comme le suggère le parallèle avec les proportions numériques, les proportions analogiques sont censées obéir à trois postulats :

- $a : b : a : b$ (*réflexivité*)
- $a : b :: c : d \Rightarrow c : d :: a : b$ (*symétrie*)
- $a : b :: c : d \Rightarrow a : c :: b : d$ (*permutation centrale*)

Si les deux premiers postulats ne sont pas surprenants, la stabilité par permutation centrale semble inattendue. Ce-

a	b	c	d
0	0	0	0
1	1	1	1
0	0	1	1
1	1	0	0
0	1	0	1
1	0	1	0

TABLE 1 – Table de vérité de $a : b :: c : d$

pendant, si l'on considère l'exemple suivant : « les chiots sont aux chiens ce que les chatons sont aux chats », il faut admettre que l'on peut aussi affirmer que « les chiots sont aux chatons ce que les chiens sont aux chats ».
L'application de la symétrie et de la permutation centrale (séquentiellement) montre qu'une proportion analogue satisfait aussi

— $a : a :: b : b$ (identité) ;
— $a : b :: c : d \Rightarrow d : b :: c : a$ (permutation externe) ;
— $a : b :: c : d \Rightarrow b : a :: d : c$ (renversement interne) ;
— $a : b :: c : d \Rightarrow d : c :: b : a$ (renversement complet).

On peut facilement vérifier que les propriétés ci-dessus s'appliquent à l'exemple « chats / chiens ».
Les entités que nous considérons sont des vecteurs de valeurs booléennes. Ainsi, chaque entité est représentée par un vecteur $\vec{a} = (a_1, ..., a_n)$ où a_i est la valeur booléenne de la caractéristique/de l'attribut i. Les proportions analogiques entre les vecteurs (de même taille) sont définies composante par composante :
$\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ si et seulement si $\forall i \in \{1, ..., n\}, a_i : b_i :: c_i : d_i$.
Nous considérons d'abord le cas d'une seule composante et de quatre variables booléennes représentant la valeur d'une caractéristique d'intérêt pour quatre éléments distincts.
Le postulat de réflexivité $a : b :: a : b$ impose qu'une proportion analogique booléenne soit vraie pour tous les a et tous les b, ce qui implique que les quatre valuations $(0, 0, 0, 0)$, $(0, 1, 0, 1)$, $(1, 0, 1, 0)$ et $(1, 1, 1, 1)$ doivent correspondre à des proportions analogiques vraies. En vertu du postulat de permutation centrale, le modèle booléen minimal qui satisfait les trois postulats est vrai pour les six configurations suivantes des variables booléennes a, b, c, d (et faux pour les $2^4 - 6 = 10$ autres configurations possibles) présentées dans la Table 1 [47].
La formule logique quaternaire suivante n'est vraie que pour les 6 configurations de valeurs ci-dessus (et donc fausse pour les 10 autres) [43] :

$$a : b :: c : d = ((a \wedge \neg b) \equiv (c \wedge \neg d)) \wedge ((\neg a \wedge b) \equiv (\neg c \wedge d)) \quad (1)$$

Cela exprime précisément que « a diffère de b comme c diffère de d, et b diffère de a comme d diffère de c » (et « lorsque a et b ne diffèrent pas, c et d ne diffèrent pas »).
Prenons l'exemple du chat et du chien : « un chiot est au chien ce qu'un chaton est au chat » pour illustrer la proportion analogique entre les vecteurs qui représentent les quatre animaux sur un ensemble de 6 attributs, ici `mammifère`, `oiseau`, `canidé`, `félin`, `jeune`, `adulte`, comme dans la Table 2.

	mam.	*ois.*	*canid.*	*félin*	*jeune*	*adul.*
chiot	1	0	1	0	1	0
chien	1	0	1	0	0	1
chaton	1	0	0	1	1	0
chat	1	0	0	1	0	1

TABLE 2 – $chiot : chien :: chaton : chat$: une validation booléenne

Comme on peut le constater, les attributs de l'exemple sont tels que chacun d'eux, verticalement, illustre l'un des 6 modèles de la Table 1. Cela confirme que la proportion analogique $chiot : chien :: chaton : chat$ est respectée.

3.2 Inférence analogique

Étant donné un triplet $a, b, c \in \{0, 1\}^3$, on peut chercher $x \in \{0, 1\}$ tel que $a : b :: c : x$ soit vrai. D'après la Table de vérité 1, on peut vérifier que si une solution x existe, elle est unique, mais il existe deux cas où l'équation n'a pas de solution. En effet, il n'existe aucun $x \in \{0, 1\}$ tel que $1 : 0 :: 0 : x$ et $0 : 1 :: 1 : x$ soient des proportions analogiques vraies. On peut vérifier que l'équation analogique booléenne $a : b :: c : x$ est résoluble si et seulement si $(a \equiv b) \vee (a \equiv c)$ est vrai, où $\equiv$ est le connecteur d'équivalence. Dans ce cas, la solution *unique* est donnée par $x = a \equiv (b \equiv c)$ (rappelons que $\equiv$ est associatif) ; ainsi $x = b$ si $a = c$ et $x = c$ si $a = b$.
L'inférence par proportions analogiques, appliquée à la classification, repose sur un principe simple : si quatre vecteurs (booléens) $\vec{a}$, $\vec{b}$, $\vec{c}$ et $\vec{d}$ forment une proportion analogique valide composante par composante pour chaque quadruplet de valeurs relatives au même attribut, alors on s'attend à ce que leurs étiquettes de classe forment également une proportion valide[3] (par exemple, [10]).

$$\frac{\vec{a} : \vec{b} :: \vec{c} : \vec{d}}{cl(\vec{a}) : cl(\vec{b}) :: cl(\vec{c}) : cl(\vec{d})} \quad (2)$$

Si les étiquettes de classe des vecteurs $\vec{a}$, $\vec{b}$ et $\vec{c}$ sont connues (c'est-à-dire qu'ils appartiennent à l'ensemble d'exemples), la classification d'un nouveau vecteur booléen $\vec{d}$ est possible à l'aide de (2) uniquement si : i) l'équation $cl(\vec{a}) : cl(\vec{b}) :: cl(\vec{c}) : x$ est résoluble ; et ii) la proportion analogue $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ est vérifiée sur toutes les composantes. Si ces deux conditions sont remplies, $cl(\vec{d})$ est la solution unique de $cl(\vec{a}) : cl(\vec{b}) :: cl(\vec{c}) : x$. En pratique, il existe généralement plusieurs triplets $(\vec{a}, \vec{b}, \vec{c})$, ce qui conduit à plusieurs solutions pour $cl(\vec{d})$, qui peuvent être différentes. Un vote majoritaire détermine alors la classe attribuée à $\vec{d}$. Cette procédure donne de bons résultats sur des benchmarks [10].

Dans [3], un schéma d'inférence légèrement différent a été proposé à des fins de créativité. À partir de trois éléments

3. De même que pour les valeurs d'attributs booléens, les seules proportions analogiques valides entre deux étiquettes de classe, éventuellement identiques, s et t, sont $s : t :: s : t$ et $s : s :: t : t$, tandis que $s : t :: t : x$ n'a pas de solution (sauf pour $s = t = x$).

$\vec{a}$, $\vec{b}$, $\vec{c}$, on peut construire un autre élément $\vec{d}$, à condition que pour chaque attribut i, l'équation booléenne analogique $a_i : b_i :: c_i : x_i$ soit résoluble, ce qui requiert que

$$(\vec{a} \equiv \vec{b}) \vee (\vec{a} \equiv \vec{c})$$

se vérifie composante par composante. Ceci peut se résumer par le schéma d'inférence suivant.

$$\frac{\vec{a}, \vec{b}, \vec{c}}{\vec{d} = (\vec{a} \equiv \vec{b} \equiv \vec{c})} \quad (3)$$

On peut démontrer que $\vec{d}$ est différent de $\vec{a}, \vec{b}, \vec{c}$, pourvu que $\vec{a} \neq \vec{b}$ et $\vec{a} \neq \vec{c}$. La section suivante étudie de combien $\vec{d}$ peut différer des trois autres vecteurs.

3.3 Une évaluation théorique de la créativité basée sur les proportions analogiques

Une proportion analogique $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ établit un parallèle entre les paires $(\vec{a}, \vec{b})$ et $(\vec{c}, \vec{d})$. Les 6 configurations du Tableau 1 qui rendent cette proportion analogique vraie nous permettent de partitionner les n attributs sur lesquels les éléments sont décrits ($\vec{a} = (a_1, ..., a_n)$, $\vec{b} = (b_1, ..., b_n)$, $\vec{c} = (c_1, ..., c_n)$, $\vec{d} = (d_1, ..., d_n)$) en 3 sous-ensembles :

- i) le sous-ensemble d'attributs i où $a_i = b_i = c_i = d_i = 1$ ou $a_i = b_i = c_i = d_i = 0$, c'est-à-dire où les quatre vecteurs $\vec{a}, \vec{b}, \vec{c}, \vec{d}$ sont égaux ;
- ii) le sous-ensemble d'attributs j où $a_j = b_j$ et $c_j = d_j$ mais $a_j \neq c_j$, c'est-à-dire où les vecteurs sont égaux au sein des paires $(\vec{a}, \vec{b})$ et $(\vec{c}, \vec{d})$, mais pas de la même manière ;
- iii) le sous-ensemble d'attribut(s) k où $a_k = c_k$ et $b_k = d_k$, c'est-à-dire le sous-ensemble d'attributs dont la ou les valeurs changent, dans la même direction, de $\vec{a}$ à $\vec{b}$ et de $\vec{c}$ à $\vec{d}$ (les différences entre $\vec{a}$ et $\vec{b}$ et entre $\vec{c}$ et entre $\vec{d}$ sont identiques).

Ceci est illustré en Table 3.

vecteurs	Tous égaux		Egaux par paires		Diff. identiq.	
$\vec{a}$	1	0	1	0	1	0
$\vec{b}$	1	0	1	0	0	1
$\vec{c}$	1	0	0	1	1	0
$\vec{d}$	1	0	0	1	0	1

TABLE 3 – Les 3 types de motifs dans une proportion analogique

Notons que la permutation centrale de $\vec{b}$ et $\vec{c}$ échange les deux dernières colonnes de la Table 3. Cela permet d'énoncer le résultat :

$\vec{a} : \vec{b} :: \vec{c} : \vec{d} \Leftrightarrow$

$$\begin{cases} Equ(\vec{a}, \vec{b}) = Equ(\vec{c}, \vec{d}) \\ Dif^{10}(\vec{a}, \vec{b}) = Dif^{10}(\vec{c}, \vec{d}) \\ Dif^{01}(\vec{a}, \vec{b}) = Dif^{01}(\vec{c}, \vec{d}) \end{cases} \Leftrightarrow \begin{cases} Equ(\vec{a}, \vec{c}) = Equ(\vec{b}, \vec{d}) \\ Dif^{10}(\vec{a}, \vec{c}) = Dif^{10}(\vec{b}, \vec{d}) \\ Dif^{01}(\vec{a}, \vec{c}) = Dif^{01}(\vec{b}, \vec{d}) \end{cases}$$

où, étant donnés deux vecteurs $\vec{x} = (x_1, ..., x_n)$ et $\vec{y} = (y_1, ..., y_n)$, on pose

$Equ(\vec{x}, \vec{y}) = \{i \mid x_i = y_i\} = Equ^0(\vec{x}, \vec{y}) \cup Equ^1(\vec{x}, \vec{y})$,

$Equ^0(\vec{x}, \vec{b}) = \{i \mid x_i = y_i = 0\}$,

$Equ^1(\vec{x}, \vec{y}) = \{i \mid x_i = y_i = 1\}$,

$Dif^{10}(\vec{x}, \vec{y}) = \{i \mid x_i = 1, y_i = 0\}$,

$Dif^{01}(\vec{x}, \vec{y}) = \{i \mid x_i = 0, y_i = 1\}$.

La distance de Hamming entre $\vec{a}$ et $\vec{b}$, $Ham(\vec{a}, \vec{b})$, est égale à la cardinalité de $Dif^{10}(\vec{a}, \vec{b}) \cup Dif^{01}(\vec{a}, \vec{b}) = Dif(\vec{a}, \vec{b})$. Nous avons $Ham(\vec{a}, \vec{b}) = |Dif(\vec{a}, \vec{b})| \triangleq e$. Notons que $Ham(\vec{a}, \vec{b}) = Ham(\vec{c}, \vec{d})$. Soit $r =$ $|\{i$ t.q. $a_i = b_i = c_i = d_i = 1$ ou $a_i = b_i = c_i = d_i = 0\}|$.

Ainsi $Ham(\vec{c}, \vec{d}) = e$; $Ham(\vec{b}, \vec{d}) = n - e - r$; $Ham(\vec{a}, \vec{d}) = n - r$.

Si $e = 0$, $\vec{a} = \vec{b}$ et $\vec{c} = \vec{d}$; si $r + e = n$, $\vec{a} = \vec{c}$, $\vec{b} = \vec{d}$. Dans les deux cas, la proportion analogique $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ est triviale : on obtient $\vec{a} : \vec{a} :: \vec{b} : \vec{b}$ ou $\vec{a} : \vec{b} :: \vec{a} : \vec{b}$. Si $r = n$, Les 4 items sont égaux.

Ainsi, $\vec{d}$ est d'autant plus éloigné de $\vec{a}$ et $\vec{c}$ que r est petit et e est grand.

Mais si $r = 0$ et $e = n$, on obtient $Ham(\vec{b}, \vec{d}) = 0$. *Il faut donc équilibrer de manière égale (ou presque égale, si n est impair) e et $n - e$, avec $r = 0$, si l'on veut maximiser la distance de $\vec{d}$ à $\{\vec{a}, \vec{b}, \vec{c}\}$* égale à $\min(e, n - e - r, n - r)$.

Si n est pair, la distance maximale est simplement $n/2$.

Le vecteur $\vec{d}$, entièrement obtenu à partir de $\vec{a}, \vec{b}, \vec{c}$, peut donc être très différent de ces trois éléments, et par conséquent tout à fait nouveau par rapport à eux.

Ainsi, l'inférence analogique (3) apparaît comme une méthode pour explorer de manière contrôlée un univers de possibilités potentielles à partir des options existantes. Cette méthode est préférable à une sélection aléatoire des valeurs des attributs (comme proposé initialement dans [37]).

4 Cadre de l'expérimentation

Nous présentons ci-dessous une expérience illustrant l'efficacité du pouvoir créatif de l'inférence par analogie basée sur les proportions. Ce processus se déroule en deux étapes : la génération de prompts, puis la production de textes. La nouveauté est vérifiée tant au niveau des prompts qu'au niveau des textes. Nous présentons d'abord notre source d'inspiration pour cette expérience, avant d'expliquer comment l'inférence par analogie basée sur les proportions est utilisée pour créer les prompts, lesquels servent ensuite à générer les textes.

La source d'inspiration L'inspiration provient d'un livre célèbre de Raymond Queneau, *Exercices de style* [49], où l'auteur réécrit 99 fois la même histoire : un court texte où le narrateur monte dans un bus, assiste à une dispute entre un homme et un autre passager, puis revoit ce même homme deux heures plus tard, en train de demander conseil sur l'ajout d'un bouton à son pardessus. À chaque fois, l'histoire est réécrite dans un style différent.[4] On peut y voir un exemple de créativité exploratoire (répétée) au sens où l'entend Boden.

Queneau est également célèbre pour un autre ouvrage [50], publié ultérieurement, qui illustre la créativité combina-

4. Il est à noter que des travaux récents [35] ont analysé, avec des embeddings, le style des *Exercices de style* (et aussi celui des *Nouvelles en trois lignes* de Félix Fénéon [29]) en le comparant à d'autres, montrant que le style rédactionnel influe sur la dispersion des embeddings, même si la variation thématique a un effet plus marqué.

toire. Ce livre est basé sur dix sonnets (poèmes de quatorze vers) dont chaque vers est imprimé sur une bande de papier épais. Les dix sonnets ont le même schéma de rimes et les mêmes sonorités. Ainsi, n'importe quel vers d'un sonnet peut être combiné avec un vers des neuf autres, permettant de créer 10^{14} poèmes différents. Ceci a conduit Queneau, avec le mathématicien François Le Lionnais, à fonder l'« Ouvroir de littérature potentielle » (Oulipo), réunissant un groupe de personnes intéressées par la littérature expérimentale.

Création des prompts Queneau, dans son ouvrage *Exercices de style* [49], varie le style à chaque réécriture de son récit, employant une grande variété de procédés rhétoriques. Pour notre expérience, nous transposons cette idée en choisissant un nombre limité de caractéristiques applicables à un texte. Nous utilisons 10 caractéristiques booléennes correspondant à la présence ou l'absence des 10 émotions suivantes : joie, peur, colère, tranquillité, nostalgie, honte, confusion, curiosité, amour et résignation.[5] Ceci présente l'avantage de fournir des vecteurs qui : i) peuvent être traités par inférence analogique ; et ii) peuvent être associés à des prompts envoyés à un chatbot basé sur un LLM capable de générer un texte contraint.

Pour constituer un *ensemble inspirant* (représentant 10% de l'ensemble des vecteurs possibles), on construit un ensemble empirique initial de 100 vecteurs (à 10 composantes) pour notre expérience ; pour ce faire nous partons d'un ensemble de 11 vecteurs : le vecteur $(0,0,0,0,0,0,0,0,0,0)$ et les 10 vecteurs comportant un seul '1' (et 9 '0') correspondant à chaque caractéristique. Il existe $2^{10} = 1024$ vecteurs booléens distincts à 10 composantes. Afin d'obtenir les $100 - 11 = 89$ vecteurs manquants de l'ensemble initial, nous avons procédé comme suit. Pour compléter l'ensemble initial, nous ajoutons les vecteurs un par un à l'ensemble de départ. A chaque étape, à l'ensemble de vecteurs courant, on associe une distribution de probabilité décrivant la probabilité d'obtenir n'importe quelle combinaison de valeurs de caractéristiques en utilisant cet ensemble comme échantillon. Le principe consiste à utiliser la distribution de probabilité de l'ensemble courant et à ajouter un vecteur (n'appartenant pas encore à l'ensemble) de manière à minimiser l'entropie de Shannon de l'ensemble résultant. Cette procédure s'arrête lorsque l'ensemble courant contient 100 vecteurs. Cet ensemble joue le rôle de l'ensemble inspirant. Appelons-le $\mathcal{I}$. Notons que $\mathcal{I}$ est relativement « homogène », en ce sens qu'il regroupe les vecteurs contenant peu de '1'.

À partir de l'ensemble $\mathcal{I}$, nous générons un ensemble de sortie $\mathcal{O}$ par inférence analogique. Pour ce faire, nous utilisons des triplets de vecteurs $(\vec{a}, \vec{b}, \vec{c})$ appartenant à $\mathcal{I}^3$ afin d'appliquer le schéma d'inférence (3). Nous appliquons une procédure semi-guidée en sélectionnant la paire candidate $(\vec{a}, \vec{b})$ qui servira ensuite à former les triplets, comme expliqué dans le paragraphe suivant. Ensuite, nous générons toutes les sorties possibles $\vec{d}$ pouvant être obtenues à partir d'une paire sélectionnée et d'un vecteur compatible $\vec{c}$ (c'est-à-dire un vecteur $\vec{c}$ pour lequel l'équation de proportion $\vec{a} : \vec{b} :: \vec{c} : x$ admet une solution), en appliquant l'inférence analogique (3). Ceci constitue l'ensemble $\mathcal{O}$.

La sélection des paires $(\vec{a}, \vec{b})$ garantit que le nombre de '1' dans $\vec{b}$ augmente par rapport à $\vec{a}$: cela signifie que $\forall i, a_i = 1 \Rightarrow b_i = 1$ et $\exists j, a_j = 0$ et $b_j = 1$. L'idée est que si une propriété est acquise lors du passage de $\vec{a}$ à $\vec{b}$, elle le sera également lors du passage de $\vec{c}$ à $\vec{d}$ (puisque 01 ne peut être complété que par 01 pour obtenir une proportion valide). De plus, le motif 1010, correspondant pour $\vec{b}$ et $\vec{d}$ à la perte répétée d'une propriété, est alors interdit. Cela ne signifie toutefois pas que $\vec{d}$ hérite de toutes les propriétés de $\vec{b}$, puisque 11 peut être complété par 11 mais aussi par 00 pour obtenir un motif de proportion analogique valide. Ce dernier point montre que le mécanisme analogique n'est pas une simple procédure d'enrichissement de propriétés, mais qu'il parvient également à construire un ensemble renouvelé de résultats à partir de l'ensemble initial, même si nous privilégions certaines paires et interdisons l'utilisation du motif 1010 dans l'inférence. En définitive, nous privilégions un ensemble de résultats constitué de nouveaux vecteurs enrichis, distincts et indépendants de l'ensemble initial.

Partant d'un ensemble $\mathcal{I}$ de taille 100, l'inférence analogique conduit à un ensemble $\mathcal{O}$ de taille approximativement 150. Chacun de ces deux ensembles peut être associé à une distribution de probabilité. Nous sommes alors en mesure de calculer la dissimilarité, ou divergence, entre les distributions $p_{\mathcal{I}}$ et $p_{\mathcal{O}}$ associées respectivement à $\mathcal{I}$ et $\mathcal{O}$.

Exploitation des prompts Les vecteurs des ensembles $\mathcal{I}$ et $\mathcal{O}$ servent de base à la construction de prompts destinés à être traités par un modèle de langage étendu (LLM) afin de générer un texte associé à chaque vecteur. En raison de l'existence d'une correspondance entre les éléments des vecteurs et la présence ou l'absence d'une émotion, la construction des prompts s'effectue en préétablissant un texte à l'avance, comportant dix variables à instancier. Ces variables sont remplacées par les chaînes de caractères « `true` » ou « `false` » en fonction de la valeur binaire de l'élément correspondant. Un exemple de prompt est présenté à la page 8. La soumission du prompt à un modèle de langage génère par la suite un texte associé à chaque vecteur.

Ces textes peuvent ensuite être lemmatisés et des n-grammes extraits (pour des valeurs fixes de n) afin de calculer une distribution de probabilité des n-grammes pour les textes générés à partir de $\mathcal{I}$, et une autre pour ceux générés à partir de $\mathcal{O}$. Ceci permet de tester, au niveau textuel, si des nouveautés sont décelables dans les textes issus de $\mathcal{O}$ par rapport à ceux issus de $\mathcal{I}$.

Pour la génération de texte, nous avons utilisé LLaMA 3. LLaMA 3 est un modèle de langage de dernière génération développé par Meta AI, conçu pour repousser les li-

5. Ce choix peut paraître quelque peu arbitraire, même s'il recoupe largement les 16 émotions de base préconisées dans [25]. Notons qu'il s'agit ici moins des émotions ressenties par les individus que d'un texte censé communiquer certains sentiments au lecteur. Le nombre (10) a été choisi car suffisamment grand pour assurer la variété, sans pour autant être trop élevé afin de limiter le coût de calcul.

mites de la génération de texte naturel grâce à une architecture sophistiquée et un grand nombre de paramètres (8 milliards). Ce modèle fait partie de la série LLaMA (Large Language Model Meta AI), qui se distingue par sa capacité à comprendre et à générer du texte de manière cohérente et contextuellement appropriée.

À titre d'avertissement, nous tenons à souligner que notre objectif n'était en aucun cas de tester la qualité des textes produits par LLaMA 3. LLaMA 3 est utilisé ici uniquement comme outil de production de textes courts, contraints de suivre la trame narrative d'une histoire banale – librement inspirée de celle qui ouvre les *Exercices de style* de Queneau – incluant ou non une série de 10 émotions. L'objectif est de constituer deux ensembles de textes, associés respectivement aux ensembles de vecteurs de $\mathcal{I}$ et $\mathcal{O}$. Il s'agit de vérifier si la nouveauté potentielle au niveau vectoriel se reflète ou non au niveau textuel. La trame narrative est spécifiée à LLaMA 3 par l'échange suivant :

user: I will give you a story and you'll have to tell me the main route of the story in a paragraph (don't jump lines keep it all together) okay?
assistant: Sounds like fun! I'm ready to hear your story. Please go ahead and share it, and I'll do my best to summarize the main route of the story in one paragraph. Go!
user: An individual boards a public transportation vehicle. His figure blends into the mass of anonymous passengers. His gaze wanders, observing the familiar faces distantly. Suddenly, a peculiarity catches his attention: a man adorned with an unusual headgear, a felt hat of austere gray, perched majestically atop his head. And there, an irregularity, a disturbance in the usual landscape: a fly, intrepid explorer, ambles nonchalantly on the nape of this impassive man. A shiver of curiosity runs through him. A question nags at his mind: should he intervene? A decision takes shape in the meanders of his thought. Does the man, this bearer of felt, deserve his attention? Is it a mere accident or a sign from destiny? An irrepressible impulse propels him to act. He rises, approaches with determination. A brief, incisive exchange, punctuated by abrupt gestures, suffices to draw the attention of the man, but in a markedly different register. A reprimand escapes his lips, cutting, imbued with thinly veiled arrogance. And the fly? It continues its journey, indifferent to this human theater, oblivious to its involuntary role in this fleeting scene of daily life
assistant: As an individual boards a public transportation vehicle, he becomes distracted by a peculiar sight: a man wearing an unusual gray felt hat and sporting an intrepid fly on his neck. The individual's curiosity is piqued, and he begins to wonder if the situation warrants intervention. As he rises from his seat and approaches the man, he delivers a reprimanding remark laced with arrogance, drawing the attention of the hat-wearing individual in a completely different register than anticipated. Meanwhile, the fly remains oblivious to its role in this brief and fleeting drama that plays out amidst the daily routines of the anonymous passengers.

Ainsi, le prompt correspondant au vecteur initial $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ est

user: Redo the main route you mentioned before transforming it in a way that dimension Joy is false, Fear is false, Anger is false, Tranquility is false, Nostalgia is false, Shame is false, Confusion is false, Curiosity is false, Love is false, Resignation is false. (don't write any of this words in the paragraph (Joy, Fear, Anger, Tranquility, Nostalgia, Shame, Confusion, Curiosity, Love, Resignation) and don't write (True or false) either.

5 Résultats expérimentaux

Tout d'abord, la comparaison des collections de vecteurs $\mathcal{I}$ et $\mathcal{O}$ dans l'expérience a donné les résultats suivants en termes de distance de Hellinger et de divergence de Kullback-Leibler : [6]

$$H(p_{\mathcal{I}}, p_{\mathcal{O}}) = 0.8625\,; \quad KL(p_{\mathcal{I}} || p_{\mathcal{O}}) = 0.6778$$

Comme on peut le constater, la nouveauté de l'ensemble de sortie $\mathcal{O}$ par rapport à l'ensemble d'origine $\mathcal{I}$ est clairement mise en évidence par la valeur élevée de la distance de Hellinger (dont le maximum possible est 1). Ceci est confirmé par l'évaluation de la divergence. Il convient toutefois de noter qu'ici, les vecteurs $\vec{d}$ de $\mathcal{O}$ ne maximisent pas nécessairement leur différence avec les vecteurs $\vec{a}, \vec{b}, \vec{c}$ de $\mathcal{I}$ dont ils sont issus, au sens de la sous-section 3.3.

Le texte produit par LLaMA 3 à partir du prompt associé au vecteur initial $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ est

assistant: As the individual steps onto the public transportation vehicle, his gaze drifts across the familiar faces, but one peculiarity stands out: a man

6. Le code et les données sont disponibles ici : https://github.com/Sasufox/AnalogyStories

`clad in an austere gray felt hat, with an intrepid explorer of a different kind - a fly - perched nonchalantly on the nape of his neck. The individual's thoughts begin to weave a tapestry of inquiry, and he rises from his seat with purpose, approaching the hat-wearer with determination. A brief, incisive exchange ensues, punctuated by swift gestures, as the two individuals intersect in a fleeting dance of daily life.`

Ce texte est assez court et dépourvu d'émotion particulière. À titre d'exemple, voici le texte produit lorsque les émotions Confusion, Curiosité et Résignation sont requises (le prompt correspond au vecteur $(0, 0, 0, 0, 0, 0, 1, 1, 0, 1)$) :

`As he boards the public transportation vehicle, his gaze drifts across the sea of anonymous faces, his mind a maze of contradictions. A peculiar sight captures his attention: a man's austere gray hat, a fly's leisurely stroll on the nape of his neck. His thoughts twist and turn, unsure whether to intervene or simply observe. The air is thick with uncertainty as he approaches the man, his words laced with an unexpected tone that belies the calm exterior. Meanwhile, the fly continues its aimless flight, seemingly unaware of the complex emotions swirling around it, leaving only a lingering sense of disorientation in its wake.`

Comme on peut le constater, la rédaction du nouveau texte a été transformée, tout en conservant la même trame narrative. Conformément aux attentes, l'ensemble des textes associés à $\mathcal{O}$ diffère de celui associé à $\mathcal{I}$. Ceci peut être vérifié à l'aide de la distance de Hellinger (qui a l'avantage d'être bornée) entre les distributions de probabilité des n-grammes dans les deux ensembles. Les résultats obtenus pour $n = 2$, $n = 3$ et $n = 4$ sont les suivants : $H(p_{\mathcal{I}}^{2-gram}, p_{\mathcal{O}}^{2-gram}) = 0.3947$: $H(p_{\mathcal{I}}^{3-gram}, p_{\mathcal{O}}^{3-gram}) = 0.6279$; $H(p_{\mathcal{I}}^{4-gram}, p_{\mathcal{O}}^{4-gram}) = 0.7690$.

Ceci confirme la nouveauté des textes de $\mathcal{O}$ par rapport à ceux de $\mathcal{I}$. Cependant, il a été noté [14] que les modèles de langage de grande taille avancés obtiennent d'excellents résultats sur des tâches d'associations sémantiques divergentes, ce qui suggère une capacité de créativité. La production de textes par les modèles de langage de grande taille étant basée sur un mécanisme aléatoire, on peut mesurer la distance entre les récits générés à partir d'un même vecteur de la manière suivante. À partir d'un échantillon représentatif (de taille 280) des 1024 vecteurs, 100 textes sont générés pour chaque vecteur. Pour chacune de ces collections, un texte est choisi comme référence et comparé aux autres textes. La distance de Hellinger entre la collection de textes de référence et la collection des autres textes peut alors être calculée. Les résultats suggèrent que LLaMA génère un grand nombre de n-grammes qui ne sont pas présents simultanément dans deux documents générés à partir du même vecteur. Ainsi, la nature aléatoire du LLM rend difficile de déterminer précisément la part de nouveauté apportée par l'utilisation d'une nouvelle consigne.

6 Discussion finale

Dans l'introduction, nous avons distingué trois types de raisonnement analogique : i) celui, parfois appelé « fusion conceptuelle », qui relie deux situations appartenant à des univers différents, comme dans l'analogie entre l'atome et le système solaire ; ii) le raisonnement à partir de cas, qui extrapole à une nouvelle situation ce qui a été observé dans un cas similaire ; iii) la résolution d'équations de proportions analogiques, qui produit un quatrième élément à partir de trois autres, comme dans les puzzles analogiques. Le premier type correspond à l'analogie transformationnelle ; le deuxième peut à peine être qualifié de créatif puisqu'il s'agit essentiellement d'adapter quelque chose de déjà connu ; le dernier type est une forme d'analogie exploratoire, où un univers (celui qui peut être décrit au moyen des attributs utilisés pour représenter les éléments considérés) est exploré.

Contrairement à un parcours purement aléatoire dans l'univers des caractéristiques booléennes, cet univers peut être exploré de manière contrôlée grâce à l'inférence basée sur les proportions analogiques, par exemple en favorisant l'accumulation de propriétés pertinentes. De manière générale, les proportions analogiques $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ garantissent que les différences entre $\vec{a}$ et $\vec{b}$ sont préservées entre $\vec{c}$ et $\vec{d}$. L'expérience a montré que même si l'ensemble d'inspiration (où $\vec{a}, \vec{b}, \vec{c}$ sont pris) est composé d'éléments relativement proches, la nouveauté de $\vec{d}$ peut être établie en termes de distance de Hellinger. Si les vecteurs décrivaient des options à un méta-niveau, le mécanisme d'inférence pourrait être utilisé pour une créativité transformationnelle (considérée comme une créativité exploratoire à ce méta-niveau).

Dans notre expérience, nous avons utilisé la minimisation de l'entropie de Shannon afin de construire un ensemble inspirant suffisamment homogène. Cela n'empêche pas le mécanisme d'inférence analogique de produire des vecteurs inédits, clairement extérieurs à cet ensemble. Il est évident que le choix de l'ensemble inspirant influence les résultats obtenus. L'utilisation de procédures de sélection en complément du mécanisme d'inférence, afin de démarrer avec des éléments « inspirants », constitue une piste importante à explorer dans de futures recherches. D'autres contraintes pourraient être envisagées pour la construction de cet ensemble.

L'expérience a démontré comment les vecteurs générés peuvent servir de base à des requêtes adressées à un chatbot basé sur un LLM, aboutissant à la production de textes « contraints » où la nouveauté inhérente au vecteur généré est partiellement préservée. Comme suggéré précédemment dans [3], l'inférence par proportions analogiques peut être utilisée pour proposer de nouveaux éléments. La combinai-

son du raisonnement analogique et des capacités de production textuelle basées sur un LLM constitue un outil de création.

Les *Exercices de style* de Queneau offrent un excellent exemple de créativité exploratoire, tout comme ses *Cent mille milliards de poèmes* [50] illustrent parfaitement l'idée de créativité combinatoire dans le domaine littéraire. Cependant, la remarquable créativité de Queneau réside dans la conception même de ces deux ouvrages, c'est-à-dire dans le dispositif lui-même. Il est intéressant de noter que la créativité de Queneau dans les *Exercices de style* est peut-être discutable, puisque Desiderius Erasmus, dans son manuel de rhétorique *De Copia* de 1512, donnait des instructions sur la manière de varier l'écriture. Ainsi, au chapitre 33, il propose 195 variations sur la même phrase (latine) signifiant « Votre lettre m'a fait très plaisir ». Queneau s'est également beaucoup intéressé à l'un des premiers recueils publiés de poésie informatique [6], où la créativité est combinatoire et l'association de mots, appartenant à un vocabulaire restreint, aléatoire. Ses commentaires, rassemblés à la fin de l'ouvrage parmi de nombreuses autres réactions d'écrivains et de linguistes, en témoignent. Il s'interroge notamment sur la possibilité pour les humains de distinguer les phrases produites par la machine de celles écrites par des poètes surréalistes, par exemple – une forme de test de Turing !

7 Conclusion

Cet article s'est concentré sur la capacité créative de l'inférence par proportions analogiques. Il apporte plusieurs contributions : premièrement, nous avons mis en évidence et illustré le potentiel créatif de cette inférence. Deuxièmement, l'évaluation de la nouveauté des résultats a été confirmée expérimentalement par les mesures de divergence appliquées aux prompts générées. La distance de Hellinger, dont les valeurs se situent dans un intervalle borné, semble particulièrement appropriée. De plus, l'expérience a également démontré la possibilité de combiner l'inférence analogique et les chatbots basés sur les LLM, tout en préservant le potentiel créatif.

Par ailleurs, plusieurs pistes de recherche méritent d'être explorées. Premièrement, le mécanisme analogique utilisé repose uniquement sur des proportions booléennes. Il pourrait être intéressant d'utiliser des proportions analogiques graduées étendues à des données numériques [24] ou la logique floue [16], par exemple pour introduire une représentation graduée des émotions dans notre expérience. De plus, dans cet article, nous nous sommes concentrés sur l'évaluation de la nouveauté des productions, mais on pourrait également vérifier la qualité des textes obtenus en estimant, à l'aide d'un lexique approprié, si ces textes contiennent effectivement des mots liés aux émotions requises par les consignes [44], ou des représentations appropriées basées sur les émotions [48]. Enfin, il peut être intéressant de comparer les résultats obtenus avec des prompts générés de manière analogique et ceux obtenus en générant directement par LLM des textes à partir de triplets d'exemples. En l'absence d'expérience directe, toute spéculation quant aux résultats qu'on obtiendrait demeure hypothétique, d'autant plus que le triplet d'exemples employé peut influencer significativement le résultat. Il est toutefois à noter que notre méthodologie contraint le modèle de langage de grande taille à exploiter l'ensemble des termes situés dans une région spécifique de l'espace de plongements associée à une émotion, tout en augmentant la diversité comme attesté par l'augmentation de la distance de Hellinger.

Le processus expérimental décrit dans cet article doit être vu comme un outil d'aide à la création plutôt que comme une machinerie créative par elle-même. En effet, de même que la créativité de Queneau dans les *Exercices de style* réside d'abord dans l'idée de produire de multiples versions d'un texte banal, jouant le rôle d'une contrainte de référence, toute en utilisant en même temps des contraintes de style différentes, ici l'utilisateur décide des paramètres à faire varier dans les textes à produire, il utilise ensuite l'inférence analogique pour explorer l'espace des possibles pour déterminer la contrainte spécifiant les paramètres qui induiront le prompt à appliquer au LLM.

Remerciements

Cette recherche a été financée par le projet de l'ANR "Analogies : from Theory to Tools and Applications" (AT2TA), ANR-22-CE23-0023.

Références

[1] Agnar Aamodt and Enric Plaza. Case-based reasoning ; foundational issues, methodological variations, and system approaches. *AICom, 7, 39-59*, 1994.

[2] Stergos Afantenos, Leonardo Cortez Bernardes, and Henri Prade. Measuring analogical creativity - an experiment. In Hannu Toivonen, Anna Kantosalo, and Hugo Gonçalo Oliveira, editors, *Proc. 17th Int. Conf. on Computational Creativity (ICCC'26), Coimbra, June 29 - July 3*, 2026.

[3] Stergos Afantenos, Henri Prade, Gilles Richard, and Leonardo Bernardes. Analogical proportions and creativity : A preliminary study. In Kazjon Grace, Maria Teresa Llano, Pedro Martins, and Maria M. Hedblom, editors, *Proc. 15th Int. Conf. on Computational Creativity (ICCC'24), Jönköping*, pages 289–298, 2024.

[4] Nelly Barbot, Laurent Miclet, and Henri Prade. Analogy between concepts. *Artif. Intell.*, 275 :487–539, 2019.

[5] Jean Barré and Thierry Poibeau. Beyond canonicity. modeling literary change in french novelist production : Canon vs. archive. In Artjoms Sela, Fotis Jannidis, and Iza Romanowska, editors, *Proc. Computational Humanities Research Conf., Paris, Dec. 6-8*, CEUR Workshop Proc., 3558, 814-830, 2023.

[6] Jean Baudot. *La Machine à Ecrire mise en marche et programmée par Jean A. Baudot*. Editions du Jour, collection 'Les poètes du jour', Montréal, 1964. 'Le

premier recueil de vers libres rédigés par un ordinateur électronique', 96p.

[7] Meriam Bayoudh, Henri Prade, and Gilles Richard. Evaluation of analogical proportions through Kolmogorov complexity. *Knowl. Based Syst.*, 29 :20–30, 2012.

[8] Margaret A. Boden, editor. *Dimensions of Creativity*. MIT Press, 1994.

[9] Margaret A. Boden. *The Creative Mind - Myths and Mechanisms (2. ed.)*. Routledge, 2003. 1st ed., Basic Books, 1990.

[10] Myriam Bounhas, Henri Prade, and Gilles Richard. Analogy-based classifiers for nominal or numerical data. *Int. J. Approx. Reason.*, 91 :36–55, 2017.

[11] Daniel G. Brown and Dan Ventura. Ethics, aesthetics and computational creativity. In Maria M. Hedblom, Anna Aurora Kantosalo, Roberto Confalonieri, Oliver Kutz, and Tony Veale, editors, *Proc. 13th Int. Conf. on Computational Creativity (ICCC'22), Bozen-Bolzano, June 27 - July 1*, pages 150–158, 2022.

[12] Bruce Buchanan. Creativity at the meta-level. *AI Magazine*, 22 :13–28, 2001.

[13] Alan Bundy. What is the difference between real creativity and mere novelty ? *Behavioral and Brain Sciences*, 17 :533–534, 1994.

[14] Honghua Chen and Nai Ding. Probing the creativity of large language models : Can models produce divergent semantic association ? *CoRR*, abs/2310.11158, 2023. Published in Findings of EMNLP 2023.

[15] Patrick Chieppe, Penny Sweetser, and Eryn Newman. Bayesian modelling of the well-made surprise In Maria M. Hedblom, Anna Aurora Kantosalo, Roberto Confalonieri, Oliver Kutz, and Tony Veale, editors, *Proc. 13th Int. Conf. on Computational Creativity (ICCC'22), Bozen-Bolzano, Jun.27-Jul. 1*, pages 126–135, 2022.

[16] Moreno Colombo, Sara D'Onofrio, and Edy Portmann. Integration of fuzzy logic in analogical reasoning : A prototype. In *Proc. 16th IEEE Int. Conf. Intelligent Computer Communic. and Processing (ICCP'20), Cluj-Napoca, Sept. 3-5*, pages 5–11, 2020.

[17] Simon Colton, Alison Pease, and Graeme D. Ritchie. The effect of input knowledge on creativity. In *Proc. Workshop on Creative Systems : Approaches to Creativity in AI and Cognitive Science, at 4th Int. Conf. on Case-Based Reasoning, Vancouver, July 31*, 2001.

[18] Simon Colton and Geraint A. Wiggins. Computational creativity : The final frontier ? In Luc De Raedt, Christian Bessiere, Didier Dubois, Patrick Doherty, Paolo Frasconi, Fredrik Heintz, and Peter J. F. Lucas, editors, *Proc. 20th Europ. Conf. on Artificial Intelligence (ECAI'12), Montpellier, Aug. 27-31*, volume 242 of *Frontiers in Artif. Intellig.and Appl.*, pages 21–26. IOS Press, 2012.

[19] Michael Cook and Simon Colton. Automated collage generation - With more intent. In D. Ventura *et al.*, editor, *Proc. 2nd Int. Conf. on Computational Creativity, Mexico City, April 27-29*, pages 1–3, 2011.

[20] Edward de Bono. *The Use of Lateral Thinking*. Avon Books, 1971.

[21] Ramón López de Mántaras. It don't mean A thing (if it ain't got that swing). In Henri Prade, editor, *Proc. 13th Europ. Conf. on Artificial Intelligence, Brighton, Aug. 23-28*, pages 694–696. John Wiley and Sons, 1998.

[22] Roy Dreistadt. An analysis of the use of analogies and metaphors in science. *The J. of Psychology*, 68 :97–116, 1968.

[23] Roy Dreistadt. The use of analogies and incubation in obtaining insights in creative problem solving. *The J. of Psychology*, 71 :159–175, 1969.

[24] Didier Dubois, Henri Prade, and Gilles Richard. Multiple-valued extensions of analogical proportions. *Fuzzy Sets Syst.*, 292 :193–202, 2016.

[25] Paul Ekman. Basic emotions. In T. Dalgleish and M. J. Power, editors, *Handbook of Cognition and Emotion*, page 45–60. John Wiley Sons Ltd, 1999.

[26] Brian Falkenhainer, Kenneth D. Forbus, and Dedre Gentner. The structure-mapping engine : algorithm and examples. *Artif. Intell.*, 41(1) :1–63, 1989.

[27] Benjamin Fele, Jan Babic, Senja Pollak, and Martin Znidarsic. Evaluation of curriculum learning algorithms using computational creativity inspired metrics. In Maria M. Hedblom, Anna Aurora Kantosalo, Roberto Confalonieri, Oliver Kutz, and Tony Veale, editors, *Proc. 13th Int. Conf. on Computational Creativity (ICCC'22), Bozen-Bolzano, June 27 - July 1*, pages 364–373, 2022.

[28] Béatrice Fuchs, Jean Lieber, Laurent Miclet, Alain Mille, Amedeo Napoli, Henri Prade, and Gilles Richard. Case-based reasoning, analogy, and interpolation. In *A Guided Tour of Artificial Intelligence Research : Vol. I : Knowledge Representation, Reasoning and Learning*, pages 307–339. Springer, 2020.

[29] Félix Fénéon. *Oeuvres*. Gallimard, 1948. Il existe de nombreuses éditions séparées des *Nouvelles en trois lignes* comme celle publiée dans la collection de poche Libretto en 2019.

[30] Dedre Gentner. Structure-mapping : A theoretical framework for analogy. *Cognitive Science*, 7(2) :155–170, 1983.

[31] Helmar Gust, Kai-Uwe Kühnberger, and Ute Schmid. Metaphors and heuristic-driven theory projection (HDTP). *Theo. Comput. Sci.*, 354(1) :98–117, 2006.

[32] Diane F. Halpern. Analogies as a critical thinking skill. In Dale Berger, Kathy Pezdek, and William Banks, editors, *Applications of Cognitive Psychology : Problem Solving, Education, and Computing*. 75-86, Routledge, 1986.

[33] Mary Hesse. On defining analogy. *Proceedings of the Aristotelian Society*, 60 :79–100, 1959.

[34] John E. Hummel and Keith J. Holyoak. Distributed representations of structure : a theory of analogical access and mapping. *Psychological Review*, 104(3) :427–466, 1997.

[35] Benjamin Icard, Evangelia Zve, Lila Sainero, Alice Breton, and Jean-Gabriel Ganascia. Embedding style beyond topics : Analyzing dispersion effects across different language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proc. 31st Int. Conf. on Computational Linguistics (COLING'25), Abu Dhabi, Jan. 19-24*, pages 3511–3522. Association for Computational Linguistics, 2025.

[36] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49 :1295–1306, 2009.

[37] Arnold Kaufmann. L'imagination artificielle - (heuristique automatique). *Revue française d'Automatique, d'Informatique et de Recherche Opérationnelle (R.A.I.R.O.) - Recherche opérationnelle*, 3(3) :5–24, 1969.

[38] Janet L. Kolodner. *Case-Based Reasoning*. Morgan Kaufmann, 1993.

[39] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22 :79–86, 1951.

[40] Mary Lou Maher. Evaluating creativity in humans, computers, and collectively intelligent systems. In Bo T. Christensen, Tore Kristensen, and Suzan Boztepe, editors, *Proc. 1st DESIRE Network Conf. on Creativity and Innovation in Design, Aarhus, Aug. 16-17*, pages 22–28. ACM, 2010.

[41] Jon McCormack and Mark d'Inverno, editors. *Computers and Creativity*. Springer, 2012.

[42] Simon McGregor. Algorithmic information theory and novelty generation. In *Proc. 4th Int. Joint Workshop on Computational Creativity, London, June 17-19*, page 109–112, 2007.

[43] Laurent Miclet and Henri Prade. Handling analogical proportions in classical logic and fuzzy logics settings. In *Proc. 10th Eur. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'09),Verona*, pages 638–650. Springer, LNCS 5590, 2009.

[44] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Iryna Gurevych and Yusuke Miyao, editors, *Proc. 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1 : Long Papers)*, pages 174–184, Melbourne, July 2018.

[45] Tiasa Mondol and Daniel G. Brown. Incorporating algorithmic information theory into fundamental concepts of computational creativity. In Andrés Gómez de Silva Garza, Tony Veale, Wendy Aguilar, and Rafael Pérez y Pérez, editors, *Proc. 12th Int. Conf. on Computational Creativity, (ICCC'21), Mexico City, (virtual), Sept. 14-18*, pages 173–181, 2021.

[46] George Polya. *Mathematics and Plausible reasoning. Volume 1 : Induction and Analogy in Mathematics*. Princeton University Press, 1954.

[47] Henri Prade and Gilles Richard. Analogical proportions : From equality to inequality. *Int. J. Approx. Reason.*, 101 :234–254, 2018.

[48] Xiangyu Qin, Zhiyu Wu, Tingting Zhang, Yanran Li, Jian Luan, Bin Wang, Li Wang, and Jinshi Cui. BERT-ERC : fine-tuning bert is enough for emotion recognition in conversation. In *Proc. 37th AAAI Conf. on Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[49] Raymond Queneau. *Exercices de Style*. Gallimard, 1947. Exercises in Style, translation by B. Wright, Gaberbocchus Press, 1958.

[50] Raymond Queneau. *Cent Mille Milliards de Poèmes*. Gallimard, 1961. Postface : F. LeLionnais. Trad. anglaise : S. Chapman, 'A Hundred Thousand Billion Poems'.

[51] Geber L. Ramalho and Jean-Gabriel Ganascia. The role of musical memory in creativity and learning : A study of jazz performance. In Matt Smith, Alan Smaill, and Geraint A. Wiggins, editors, *Music Education : An Artificial Intelligence Approach, Proceedings of a Workshop held as part of AI-ED 93, World Conf. on Artificial Intelligence in Education, Edinburgh, 25 Aug.*, Workshops in Computing, pages 143–156. Springer, 1993.

[52] Geber L. Ramalho and Jean-Gabriel Ganascia. Simulating creativity in jazz performance. In Barbara Hayes-Roth and Richard E. Korf, editors, *Proc. 12th National Conf. on Artificial Intelligence (AAAI'94), Seattle, Jul. 31 - Aug. 4, Volume 1*, pages 108–113. AAAI Press / The MIT Press, 1994.

[53] Mel Rhodes. An analysis of creativity. *Phi Delta Kappan*, 42 :305–310, 1961.

[54] Graeme Ritchie. Assessing creativity. In *Proc. of the AISB Symp. on Artificial Intelligence and Creativity in Arts and Science, York*, pages 3–11, 2001.

[55] Graeme Ritchie. The transformational creativity hypothesis. *New Generation Computing*, 24 :241–266, 2006.

[56] Graeme Ritchie. Some empirical criteria for attributing creativity to a computer program. *Minds Mach.*, 17(1) :67–99, 2007.

[57] Graeme Ritchie. A closer look at creativity as search. In Mary Lou Maher, Kristian J. Hammond, Alison Pease, Rafael Pérez y Pérez, Dan Ventura, and Geraint A. Wiggins, editors, *Proc. 3rd Int. Conf. on Computational Creativity (ICCC'12), Dublin, May 30 - June 1*, pages 41–48, 2012.

[58] Luís Espírito Santo, Geraint A. Wiggins, and Amílcar Cardoso. Towards a formal creativity theory : Preliminary results in novelty and transformativeness. *CoRR*, abs/2405.02148, 2024.

[59] Roger Schank and Alex Kass. Explanations, machine learning, and creativity. In Yves Kodratoff and Ryszard S. Michalski, editors, *Machine Learning : an Artificial Intelligence Approach, vol. III*, page 31–59. Morgan Kaufmann Publishers, San Francisco, 1990.

[60] Jürgen Schmidhuber. Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In Vincent Corruble, Masayuki Takeda, and Einoshin Suzuki, editors, *Proc. 10th Int. Conf. on Discovery Science (DS'07), Sendai, Oct. 1-4*, LNCS, pages 26–38. Springer, 2007.

[61] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Trans. on Autonomous Mental Development*, 2(3) :230–247, 2010.

[62] Martin Schmidt, Ulf Krumnack, Helmar Gust, and Kai-Uwe Kühnberger. Heuristic-driven theory projection : An overview. In Henri Prade and Gilles Richard, editors, *Computational Approaches to Analogical Reasoning : Current Trends*, Studies in Comput. Intelligence, 548, 163-194. Springer, 2014.

[63] Mark Turner and Gilles Fauconnier. *The Way We Think : Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, 2003.

[64] Tim Van de Cruys. Artificial intelligence and literature. In Pierre Marquis, Odile Papini, and Henri Prade, editors, *A Guided Tour of Artificial Intelligence Research : Volume III : Interfaces and Applications of Artificial Intelligence*, pages 487–501. Springer, 2020.

[65] Manuela M. Veloso. Prodigy/analogy : Analogical reasoning in general problem solving. In Stefan Wess, Klaus-Dieter Althoff, and Michael M. Richter, editors, *Topics in Case-Based Reasoning, 1st Europ. Workshop (EWCBR'93), Kaiserslautern, Nov. 1-5, Selected Papers*, LNCS, 837, 33-52. Springer, 1993.

[66] Geraint A. WigginsA preliminary framework for description, analysis and comparison of creative systems. *Knowl.based Syst.*,19(7) :449–458, 2006.

[67] Geraint A. Wiggins. Searching for computational creativity. *New Gener. Comput.*, 24(3) :209–222, 2006.

[68] Geraint A. Wiggins. A framework for description, analysis and comparison of creative systems. In Tony Veale and F. Amílcar Cardoso, editors, *Computational Creativity - The Philosophy and Engineering of Autonomously Creative Systems*, pages 21–47. Springer, 2019.

[69] Geraint A. Wiggins and Jamie Forth. Computational creativity and live algorithms. In Alex McLean and Roger T. Dean, editors, *The Oxford Handbook of Algorithmic Music*. Oxford University Press, 2018.