Exploiter les présupposés implicites pour qualifier les divergences entre un texte argumentatif et un corpus de référence

Loïc Récoupé[1], Marie-Hélène Abel[1]

[1] Université de Technologie de Compiègne, CNRS, Heudiasyc
CS 60319, 60203 Compiègne Cedex, France

loic.recoupe@etu.utc.fr, marie-helene.abel@utc.fr

Résumé

Les grands modèles de langage et les systèmes RAG peinent à gérer les désaccords profonds entre documents sans lisser les tensions. Nous proposons CADE, une approche asymétrique de cartographie des conflits épistémiques. Le challenger subit une extraction argumentative descendante qui infère ses présupposés implicites, ou warrants (selon le modèle de Toulmin). Le corpus de référence reçoit un traitement différent : nous y synthétisons les positions explicites et leurs cibles d'attaque. La confrontation enchaîne deux couches livrables : une cartographie verbatim des engagements du corpus, puis une lecture épistémique qui qualifie chaque friction selon une taxonomie à trois niveaux (factuelle, mécanique, axiologique). Le système cartographie et qualifie les frictions sans trancher la vérité, dans la double relativité au corpus choisi et à la grille de lecture qui opérationnalise le framework. Nous présentons l'approche, son pipeline asymétrique en sept étapes et son schéma de données. Une instance d'évaluation préliminaire a été conduite ; sa caractérisation complète fera l'objet d'une publication étendue.

Mots-clés

fouille d'arguments, argumentation computationnelle, prémisses implicites, conflits de connaissances, grands modèles de langage.

Abstract

Large Language Models and RAG systems struggle to manage deep documentary disagreements without smoothing out tensions. We propose CADE, an asymmetric approach for mapping epistemic conflicts. The challenger undergoes a top-down argumentative extraction that infers its implicit presuppositions, or warrants (following Toulmin's model). The reference corpus receives a different treatment : we synthesize its explicit positions and their attack targets. The confrontation chains two deliverable layers : a verbatim mapping of corpus engagements, then an epistemic reading that qualifies each friction according to a three-level taxonomy (factual, mechanical, axiological). The system maps and qualifies frictions without arbitrating truth, within a double relativity to the chosen corpus and to the reading grid that operationalizes the framework. We present the approach, its seven-step asymmetric pipeline and its data schema. A preliminary evaluation instance has been conducted ; its full characterization will be developed in an extended publication.

Keywords

argument mining, computational argumentation, implicit premises, knowledge conflicts, large language models.

1 Introduction

Face à un argumentaire contradictoire, l'enjeu principal est de comprendre *sur quoi* porte le désaccord et si la divergence est résolvable. Déconstruire un raisonnement adverse, identifier ses prémisses et le confronter à ce que nous savons demande un effort considérable, rarement proportionnel au temps disponible. En pratique, nous prenons des raccourcis : nous acceptons ou rejetons une thèse sans formuler explicitement pourquoi.

Les grands modèles de langage (LLM) offrent des capacités d'analyse textuelle qui pourraient assister ce travail. Mais en pratique, le processus d'affinage par RLHF (*Reinforcement Learning from Human Feedback*) pousse les modèles vers la complaisance plutôt que vers l'explicitation des tensions [13]. Les systèmes RAG (*Retrieval-Augmented Generation*), lorsqu'ils récupèrent des documents contradictoires, s'alignent sur la source la plus fréquente sans expliciter le conflit [15]. Les graphes de connaissances de type OWL exigent la cohérence logique formelle et peinent à représenter la coexistence de perspectives contradictoires, ce que les frameworks d'argumentation abstraite [2] cherchent à pallier.

Pour répondre à ce problème, plutôt que de comparer globalement deux textes ou de chercher à établir une vérité, nous avons défini l'approche CADE (*Cartographie Assistée des Désaccords Épistémiques*). Avec cette dernière, nous proposons d'extraire les *présupposés implicites* d'un texte argumentatif (ses *warrants*, au sens de Toulmin [14]) et de les utiliser comme requêtes ciblées contre un corpus de référence. Ce corpus est pragmatiquement un proxy du système de croyances de l'utilisateur. L'objectif est de l'assister en produisant une cartographie des divergences entre le texte et ce référentiel, qualifiées selon leur profondeur : factuelle, mécanique ou axiologique.

Nous présentons d'abord le cadre théorique sur lequel repose l'approche CADE (section 2), puis les travaux sur lesquels elle s'appuie et par rapport auxquels elle se situe (section 3). La section 4 décrit l'approche elle-même et le pipeline défini. La section 5 discute les limites et les perspectives.

2 Cadre théorique

L'approche CADE repose sur la confrontation d'un texte argumentatif, que nous appellerons le *challenger*, avec une base documentaire ciblée, le *corpus de référence*. Ce corpus a vocation à incarner le système de croyances et de connaissances de l'utilisateur, afin de l'assister dans la compréhension des tensions face au texte analysé. Pour opérer cette confrontation, nous mobilisons un modèle de décomposition argumentative, une hiérarchie des niveaux de désaccord pour qualifier les frictions, et une posture épistémique qui définit les limites du système.

2.1 Le warrant comme présupposé implicite

Le modèle de Toulmin [14] décompose un argument en trois éléments. Les *données* (D) sont les prémisses factuelles avancées par l'auteur, et la *conclusion* ou *claim* (C) est l'assertion qu'il en tire. Le *warrant* (W) est la proposition qui fait le pont logique entre les deux : le présupposé nécessaire pour que l'inférence $D \rightarrow C$ tienne, souvent implicite.

Cette définition implique une frontière claire avec les présupposés énoncés explicitement dans le texte. Quand un auteur formule lui-même le principe qui autorise son inférence, ce principe entre dans la décomposition argumentative comme *claim* ou comme donnée selon son statut local, et n'a pas à être inféré. La tâche d'inférence des warrants ne s'active donc que sur les sauts inférentiels que l'auteur laisse implicites.

Considérons l'exemple suivant : un commentateur observe que le pays X obtient les meilleurs scores aux évaluations PISA (D) et conclut que son système éducatif devrait servir de modèle pour réformer le nôtre (C). Le lien semble naturel, mais il suppose implicitement que les résultats PISA sont attribuables au système éducatif lui-même (W) plutôt qu'à des facteurs externes (investissement parental, culture de la réussite scolaire, structure sociale). L'auteur ne formule jamais ce warrant car il le tient pour évident. Dans notre système, extraire un warrant revient simplement à identifier ce postulat logique nécessaire à la validité de l'inférence.

2.2 Profondeur du désaccord

Une fois les warrants du challenger extraits, le système les confronte aux documents du corpus. Lorsque le corpus contredit un warrant, il faut qualifier la nature de cette divergence. La *théorie des stases*[1] structure cette qualification autour du *point d'arrêt* du désaccord : porte-t-il sur les faits, sur le mécanisme ou sur les valeurs ?

En projetant cette approche sur la décomposition de Toulmin, nous obtenons une taxonomie à trois niveaux appliquée aux interactions entre le challenger et les sources du corpus :

- **Factuel** : le corpus conteste les données que le warrant présuppose (un chiffre, une observation). C'est le niveau le plus directement résolvable.
- **Mécanique** : les faits ne sont pas contestés, mais le lien causal ou logique l'est. Dans l'exemple PISA : le corpus ne conteste pas les scores, il conteste l'attribution causale.
- **Axiologique** : faits et mécanismes peuvent être partagés, la contestation porte sur les valeurs ou les priorités sous-jacentes. Ce niveau est rarement résolvable par les données seules.

Une précision conceptuelle s'impose ici. La nature propositionnelle d'un *warrant* et le niveau de la stase d'un désaccord sont deux choses distinctes. Un *warrant* peut être par sa formulation factuel (« X est un fait établi »), mécanique (« X cause Y ») ou axiologique (« X est désirable »). Le niveau de stase, lui, qualifie la friction qui émerge entre ce *warrant* et la position que le corpus lui oppose. Les deux ne sont pas nécessairement alignés. Un *warrant* mécanique peut être contesté axiologiquement, « oui, X cause Y, mais Y n'est pas désirable », et le système doit pouvoir capter ce cas. La classification F/M/A qualifie donc la paire *warrant*-engagement.

Le type de désaccord émerge de la *confrontation* entre le warrant et ce que le corpus lui oppose. Typer un warrant avant cette confrontation serait prématuré.

2.3 Posture du système

Le système cartographie les relations entre le document analysé et le corpus de référence. Il relève le terrain commun (les accords point par point), les angles morts (ce que le corpus ne couvre pas), et qualifie les frictions (les divergences) selon la taxonomie Factuelle / Mécanique / Axiologique (F/M/A). Ce diagnostic est toujours *relatif* au corpus choisi. Un même document confronté à deux corpus différents produira naturellement des résultats différents. Cette relativité est un paramètre intrinsèque de l'approche. L'utilisateur, en choisissant son corpus, établit sa propre perspective de référence. Une seconde relativité, structurelle, s'y superpose : la formulation des prompts qui opérationnalise le cadre constitue une grille de lecture, dont les conséquences sont discutées en section 5.

Cette posture est aussi formalisable. Gómez Álvarez et al. [3] montrent avec la *Standpoint Logic* que la coexistence de perspectives contradictoires est décidable : deux sources qui se contredisent ne sont pas une incohérence logique si leurs assertions relèvent de points de vue distincts. Le système opère dans cette logique. Ce travail est mobilisé ici comme ancrage conceptuel : la posture de non-arbitrage de CADE est cohérente avec ses propriétés formelles, sans que l'implémentation en dépende directement.

1. Développée dans l'Antiquité par Hermagoras puis Cicéron, la théorie des stases définit les points d'arrêt d'un désaccord : on ne peut débattre de l'action à mener sans s'accorder sur la nature du problème, ni convenir de la nature du problème sans s'accorder sur les faits. Voir https://icar.cnrs.fr/dicoplantin/stase/

Cette posture impose des contraintes d'intégrité textuelle fortes : aucune assertion produite par le système ne doit déformer la matière source. Le *pipeline* implémente cette contrainte par un mécanisme de *span-pointing* : le LLM ne renvoie que des identifiants de phrases numérotées, le code reconstitue le texte.

3 Travaux connexes

Notre travail se situe au croisement de trois axes de recherche : l'extraction automatique de la structure argumentative, la reconstruction des présupposés implicites, et la gestion des conflits entre sources.

Extraction argumentative. L'*argument mining* vise à extraire d'un texte ses composants argumentatifs (claims, prémisses) et les relations qui les lient [8]. Les approches récentes exploitent les LLM de différentes manières. AMELIA [11] entraîne un modèle multi-tâche unifié sur 19 corpus et introduit une taxonomie des types de preuves (*Study*, *Expert*, *Anecdotal*, *Explanation*) qui qualifie la nature des prémisses. Gorur et al. [4] montrent qu'un LLM généraliste en *few-shot* peut rivaliser avec des modèles spécialisés pour la classification de relations argumentatives, avec une meilleure robustesse au changement de genre textuel. Ces travaux traitent de l'extraction de la structure de surface. La question des présupposés implicites qui sous-tendent les inférences n'entre pas dans leur périmètre. Côté ontologies, des modèles sémantiques dédiés à la représentation du discours argumentatif (la *Document Components Ontology* (DoCO) [1], l'*Argument Model Ontology* [9] ou l'Argument Interchange Format [10]) offrent un cadre alternatif à la cohérence logique formelle des graphes OWL. Notre approche vise un objectif distinct : produire un diagnostic de friction entre un texte et un référentiel, sans construire d'ontologie argumentative réutilisable.

Warrants et présupposés implicites. La tâche d'identification des warrants a d'abord été formalisée comme un problème de sélection. La *shared task* SemEval-2018 [6] proposait de choisir le bon warrant parmi deux options pour chaque paire (prémisse, claim). Le meilleur système atteignait 0.71 ; identifier les warrants nécessite des connaissances qui ne sont pas dans le texte. Plus récemment, Gupta et al. [5] montrent que les LLM, invités à raisonner selon le modèle de Toulmin, produisent des warrants jugés acceptables dans 61,7 % des cas par des annotateurs humains (contre 45,7 % pour les annotations humaines de référence, un écart qui reflète la subjectivité inhérente à la tâche). L'approche ne génère toutefois qu'un seul warrant par paire, sur des textes courts.

Qualifier les conflits documentaires. Une fois la structure argumentative et ses présupposés extraits, se pose le défi de la confrontation à des sources divergentes. Xu et al. [15] dressent un état des lieux des conflits de connaissances dans les systèmes à base de LLM (conflits intra-mémoire, incohérences entre récupérations textuelles). Si leur travail documente le problème et ses manifestations, il ne propose pas de mécanisme pour *qualifier* la nature du conflit. Par ailleurs, les LLM peinent à gérer les désaccords sans tomber dans la complaisance : le phénomène de *sycophancy* [13] pousse le modèle à s'aligner sur l'avis implicite de l'utilisateur plutôt qu'à expliciter une tension argumentative.

Positionnement. Notre travail porte sur l'articulation de ces trois axes. L'extraction argumentative fournit la structure, les warrants fournissent les hypothèses testables, et la confrontation au corpus vise à qualifier les divergences. L'idée de transformer les présupposés implicites en requêtes ciblées pour le *retrieval*, puis de qualifier chaque friction selon la taxonomie factuel / mécanique / axiologique (section 2), est la contribution que nous explorons dans ce travail.

4 L'approche CADE

Extraire la structure argumentative, inférer les présupposés implicites, confronter ces présupposés au corpus : l'approche CADE enchaîne ces opérations pour transformer la recherche documentaire d'informations ouvertes en une vérification d'hypothèses ciblées. Concrètement, nous extrayons les *warrants* (au sens de Toulmin [14]) d'un document argumentatif et les utilisons comme requêtes contre un corpus de référence. Chaque *warrant* devient une hypothèse testable ; le corpus fournit les éléments pour la soutenir, la nuancer ou la contester, et le diagnostic en qualifie la nature.

4.1 Le pipeline asymétrique en sept étapes

Le *pipeline* s'organise en trois phases (figure 1) : analyse, distillation, confrontation. L'asymétrie est la propriété structurante du système, et elle opère à deux niveaux distincts. En distillation, nous inférons l'implicite côté *challenger* (les warrants) et synthétisons l'explicite côté corpus (positions, thèses, réfutations). En confrontation, nous ne projetons pas d'interprétation sur le corpus : nous y cherchons les formulations explicites qui engagent les présupposés du *challenger*. Ce choix est délibéré et découle du statut de chaque texte : le *challenger* est l'objet d'analyse ; le corpus est le référentiel, lu selon ses propres affirmations.

La **Phase A (Analyse)** est commune aux deux branches. Chaque document traverse les mêmes micro-tâches, gérées par un *LLM* en configuration *few-shot* : Gorur et al. [4] montrent qu'une telle approche généralise mieux d'un genre textuel à l'autre. Le document brut est structuré en sections avec leurs rôles rhétoriques (1). Chaque section est analysée pour extraire les arguments (*claims* et données), les relations internes et les positions attaquées (2). Les données sont qualifiées selon la taxonomie des types de preuves d'AMELIA [11] (*Study*, *Expert*, *Anecdotal*, *Explanation*). Cette qualification sert ici de métadonnée d'audit pour l'analyste ; elle n'est pas exploitée comme signal de calcul par les passes aval. La distinction opérationnelle entre données et *claims* repose sur leur statut argumentatif local : un *claim* est une assertion défendue par d'autres éléments du texte ; une donnée est invoquée comme appui factuel sans être elle-même justifiée (observation, chiffre, citation d'autorité). Une même proposition peut être donnée dans un texte et *claim* dans un autre, selon qu'elle est

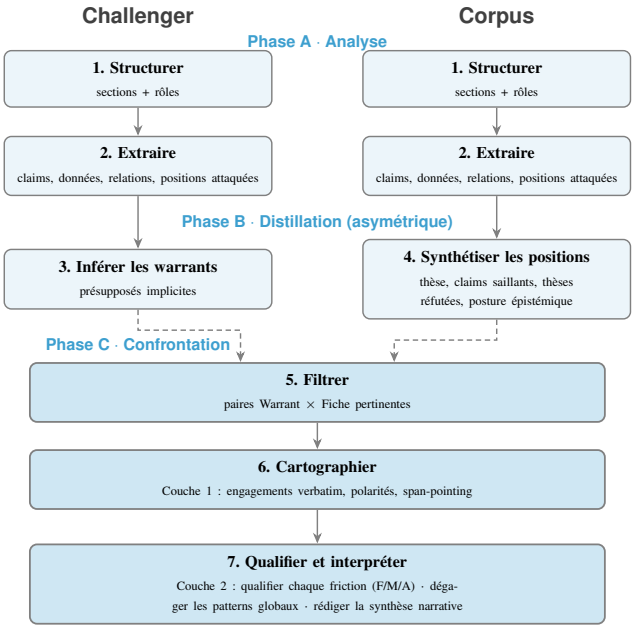

FIGURE 1 – Le pipeline asymétrique en sept étapes, organisé en trois phases. Phase A (Analyse) couvre la structuration et l'extraction argumentative communes aux deux branches. Phase B (Distillation), asymétrique, infère les présupposés implicites côté *challenger* (warrants) et synthétise les positions explicites côté corpus (fiches de position). Phase C (Confrontation) enchaîne un filtrage global et deux couches livrables : la Couche 1 (Cartographie) extrait verbatim les engagements du corpus, la Couche 2 (Lecture épistémique) qualifie chaque friction selon la taxonomie F/M/A et compose une synthèse narrative.

ou non l'objet d'une justification dans ce texte.

La **Phase B (Distillation)** est asymétrique : le traitement diverge selon le rôle du document. **Côté challenger (3) :** pour chaque relation identifiée entre une donnée et un *claim*, le système infère une multiplicité de *warrants* implicites conjonctifs. Reprenons l'illustration PISA : de l'observation des scores (donnée) à l'appel à la réforme (*claim*), la validité de l'inférence repose sur le *warrant* W_1 (PISA mesure des indicateurs fiables) et W_2 (les résultats découlent de l'organisation scolaire). La génération est *multi-warrants* : contrairement aux approches n'en isolant qu'un par lien [5], nous retenons l'hypothèse de multiples axiomes conjoints. La question posée au modèle formule les postulats logiques abstraits qui rendent la déduction valide, indépendamment de l'intention de l'auteur. Une passe de regroupement vectoriel par similarité cosinus regroupe ensuite les warrants redondants en clusters ; chaque warrant conserve sa source originale dans le graphe.

Côté corpus (4) : les étapes 1 et 2 s'appliquent identiquement. La distillation diverge ensuite. Chaque document est synthétisé en une *fiche de position* : thèse centrale reformulée de façon autoportante, *claims* explicitement défendus, thèses adverses identifiées, étiquette de posture épistémique parmi un petit ensemble de catégories opérationnelles. Nous n'inférons délibérément aucun warrant côté corpus : le référentiel est lu selon ses affirmations explicites.

Confrontation. La phase enchaîne trois étapes en deux couches livrables. (5) Un filtrage global identifie, en un appel unique, les paires (*warrant*, fiche de position) où un engagement argumentatif est plausible. Les paires sans signal de résonance sont écartées ; les warrants sans couverture dans le corpus sont signalés. Ces angles morts sont un livrable épistémique à part entière. (6) La **Couche 1** (cartographie) extrait verbatim, pour chaque paire retenue, les passages du corpus qui engagent le *warrant*, avec leur polarité (soutien, nuance, contestation) et leur traçabilité textuelle. Le signal le plus net est la correspondance directe entre un *warrant* et une thèse explicitement réfutée par le corpus : la friction est alors bidirectionnelle. (7) La **Couche 2** (lecture épistémique) comprend trois sous-passes : qualification F/M/A de chaque friction, consolidation macro (warrants centraux, angles morts, convergences paradoxales), composition d'une synthèse narrative.

Le filtrage (étape 5) opère à la granularité (*warrant*, document corpus). La cartographie (étape 6) descend à la granularité (*warrant*, point d'engagement) : un même *warrant* peut recevoir plusieurs engagements distincts d'un même document, chacun avec sa polarité propre. La qualification F/M/A opère à cette granularité.

4.2 Le schéma de données et la traçabilité

Pour garantir l'indépendance et l'intégrité de ces sept étapes, elles partagent un même contrat de données modélisé par un graphe orienté étiqueté. Ce modèle se projette sur trois couches conceptuelles.

La *couche documentaire* figure une relation de méronymie : un document contient des sections, qui contiennent des arguments caractérisés par leur texte source et leur type empirique. Cette strate est identique côté challenger et côté corpus. La *couche argumentative* formalise les dépendances causales entre arguments (soutiens et attaques). Côté challenger uniquement, des nœuds Warrant s'intercalent aux arêtes reliant une donnée à sa conclusion. La *couche de confrontation* porte les arêtes étiquetées (*Contests*, *Supports*, *Nuances*) qui relient les warrants du *challenger* aux passages explicites du corpus, avec leurs justifications et leur traçabilité textuelle.

Cet agencement a pour objectif premier la **traçabilité**. La structure trace un parcours immuable : face à une faille recensée dans le diagnostic global, l'analyste peut remonter l'arborescence jusqu'au *warrant* impliqué, identifier la chaîne d'inférence fautive du challenger, et consulter les passages d'origine, à la fois pour le document interrogé et pour les fragments justificatifs issus du corpus.

4.3 Diagnostic et restitution

Pour chaque *warrant* identifié, le *pipeline* restitue son statut global : soutenu (consensus partagé), contesté (friction décelée), nuancé (révèle une forme de complexité sans contredire abruptement) ou non couvert (angle mort propre au corpus). Rapporter les *warrants* partagés est aussi impor-

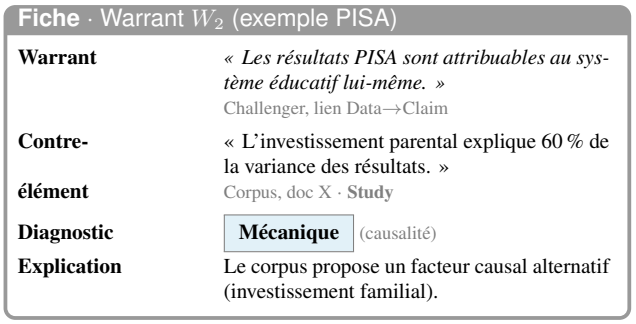

FIGURE 2 – Exemple de fiche de friction condensée (le *claim* et la donnée source ont été omis pour la lisibilité). Chaque élément est classifié et sourcé.

tant que relever les frictions, pour ne pas surreprésenter les tensions dans le diagnostic.

Concernant les *warrants* contestés, le système produit un diagnostic Factuel / Mécanique / Axiologique (F/M/A). Ce triptyque renseigne sur la profondeur du désaccord. Dans l'exemple de la figure 2, le corpus oppose une méta-analyse (*Study*) qui établit que l'investissement parental explique à 60 % le succès scolaire. Ce fragment ne conteste pas la mesure des scores PISA, mais leur attribution au seul enseignement scolaire ; la friction porte sur le lien causal du *warrant* W_2. Le désaccord est donc jugé *mécanique*. Face au même contexte, contester l'intégrité de la note PISA atteindrait W_1 par une contradiction purement *factuelle* ; refuser le modèle de l'école au nom de valeurs sociétales soulèverait un différend *axiologique*. La nature de la preuve invoquée côté corpus (ici une quantification empirique) oriente le type de désaccord identifié.

Le *pipeline* est intégralement implémenté (étapes 1 à 7). Nous en rapportons ici une instance préliminaire d'évaluation. Le *challenger* est un article d'Olivier Galland (2020) défendant la réforme des retraites comme levier de rupture avec une « société de défiance ». Le corpus de référence rassemble cinq documents : une analyse macroéconomique contestant le déficit structurel invoqué (Sterdyniak, 2023), une théorie de la retraite comme salaire à la qualification (Friot, 2023), une critique de la novlangue néolibérale (Bourdieu et Wacquant, 2000), une défense constitutionnelle de la solidarité institutionnalisée (Supiot, 2021), et un article hors périmètre thématique sur l'hôpital public, inséré comme distracteur pour tester la robustesse du filtrage (Coutinet et Domin, 2020). L'exécution mobilise le modèle `gemma-4-31B-it-UD-Q4_K_XL` (quantification GGUF, mode raisonnement natif, 24 Go VRAM) pour les 74 appels LLM du *pipeline*. L'évaluation systématique des sorties a été conduite par un évaluateur externe distinct et plus puissant, Claude Opus 4.7 en mode raisonnement maximal, selon un protocole *LLM-as-a-judge* séparé par phase. La caractérisation complète de cette instance et les tests diagnostiques associés relèvent d'une publication étendue ; nous n'en rapportons ici que les observations les plus structurantes.

Sur les 26 *warrants* extraits du *challenger*, le clustering produit 10 à 12 groupes conceptuellement distincts. Au niveau des clusters, le *pipeline* converge à 83 % avec une baseline produite indépendamment par l'évaluateur (10 des 12 présupposés implicites de la baseline retrouvés). Le résultat le plus inattendu concerne les convergences de méthode entre textes opposés. Galland défend la réforme en arguant que le choix de l'instrument (la fusion des régimes) trahit une motivation politique cachée : casser le corporatisme. Sterdyniak critique la même réforme en arguant que le choix de l'instrument (le recul de l'âge) trahit une motivation politique cachée : réduire les dépenses sociales. Les deux mobilisent le même *warrant* : le choix de l'instrument révèle l'intention authentique sous l'argumentation officielle. La cartographie par *warrants* rend ce parallèle immédiatement lisible, là où une lecture linéaire des textes ne le révèle pas. Ce résultat n'est documenté que sur un cas ; il appelle confirmation sur d'autres corpus.

L'architecture en deux couches (filtrage sur résumés, puis cartographie sur textes intégraux) régule activement les erreurs introduites en amont. Sur la run Galland, plusieurs paires classées SUPPORTS au filtrage global ont été inversées en CONTESTS lors de la cartographie verbatim, après lecture du texte intégral du document corpus. Le mécanisme est reproductible : la passe rapide sur résumés privilégie l'appariement lexical, la passe profonde rétablit la polarité effective. Ce phénomène valide l'utilité de la seconde couche comme correcteur, au-delà de la simple extraction.

Le *pipeline* identifie aussi les *warrants* centraux du *challenger* qui ne reçoivent aucun engagement du corpus, ni soutien, ni nuance, ni contestation. Cette cartographie négative est un livrable épistémique : elle montre à l'analyste où son propre référentiel est silencieux et donc insuffisant pour évaluer l'argument. Un système qui ne signalerait que les frictions sur-représenterait la conflictualité et masquerait les zones d'ignorance.

Le diagnostic n'est pas exempt de biais structurels. Sur les 14 qualifications F/M/A retenues, la distribution observée est de 1 Factuel, 12 Mécanique, 1 Axiologique. Pour un face-à-face où les enjeux axiologiques sont centraux, cette sous-représentation est marquée. L'hypothèse est double : d'une part, la décomposition de Toulmin formule les *warrants* comme principes causaux ($D \rightarrow C$), ce qui pré-cadre structurellement la lecture de la friction au niveau du mécanisme ; d'autre part, la contrainte d'abstraction du prompt pousse vers la forme causale par construction. Ce M-bias illustre comment la formulation opérationnelle d'un cadre théorique pré-cadre l'espace des observations possibles. Le départage entre attribution paradigmatique (couplage Toulmin $\times$ stases) et attribution d'implémentation (formulation du prompt) reste un test diagnostique ouvert.

5 Discussion et perspectives

5.1 Limites

La posture de non-arbitrage du système produit un diagnostic doublement relatif : au corpus de référence choisi, et à la formulation des prompts qui opérationnalise le *fra-*

mework. La première relativité est revendiquée comme paramètre intrinsèque de l'approche (cf. section 2). La seconde découle de toute opérationnalisation algorithmique d'un cadre conceptuel : la grille de lecture est inspectable et tracée, mais elle n'est pas neutre. La traduction d'une théorie en contraintes de prompt influence directement ce que le système est capable d'observer.

La décomposition locale par arête (un *warrant* par lien donnée $\rightarrow$ conclusion) produit des objets locaux par construction ; les stratégies rhétoriques qui n'existent qu'à l'échelle du document entier lui échappent structurellement. Le protocole d'évaluation comporte par ailleurs des limites assumées : un seul cas évalué, un évaluateur LLM unique, absence de calibration contre une annotation humaine *gold*. Ces contraintes bornent la portée du retour d'expérience préliminaire.

5.2 Extensions

Ce travail propose CADE comme cadre pour rendre lisible sur quoi se joue un désaccord documentaire, en transformant les présupposés implicites d'un texte en hypothèses testables contre un référentiel choisi. L'instance d'évaluation préliminaire documente des zones de fiabilité et des modes de défaillance dont la caractérisation complète relève d'une publication étendue.

Plusieurs directions étendent le cadre actuel : raisonnement multi-hop inter-documents par propagation d'activation le long d'un graphe argumentatif [7] ; intégration de sources purement factuelles via un schéma de données hétérogène ; quantification de la force des désaccords par les fonctions de croyance [12] ; reformulation adversariale multi-agents pour atténuer le biais d'agrément. Ces directions restent ouvertes.

Références

[1] Alexandru Constantin, Silvio Peroni, Steve Pettifer, David Shotton, and Fabio Vitali. The document components ontology (DoCO). *Semantic Web*, 2016.

[2] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 1995.

[3] Lucía Gómez Álvarez and Sebastian Rudolph. Standpoint logic : Multi-perspective knowledge representation. In *Proceedings of the 12th International Conference on Formal Ontology in Information Systems (FOIS 2021)*, 2021.

[4] Deniz Görur, Antonio Rago, and Francesca Toni. Can large language models perform relation-based argument mining ? In *Proceedings of the 31st International Conference on Computational Linguistics*, 2025.

[5] Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. Harnessing Toulmin's theory for zero-shot argument explication. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 2024.

[6] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Semeval-2018 task 12 : The argument reasoning comprehension task. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, 2018.

[7] Hanqi Jiang, Junhao Chen, Yi Pan, Ling Chen, Weihang You, Yifan Zhou, Ruidong Zhang, Andrea Sikora, Lin Zhao, Yohannes Abate, and Tianming Liu. SYNAPSE : Empowering LLM agents with episodic-semantic memory via spreading activation. *arXiv preprint arXiv :2601.02744*, 2026.

[8] John Lawrence and Chris Reed. Argument mining : A survey. *Computational Linguistics*, 2019.

[9] Silvio Peroni and Fabio Vitali. The argument model ontology (AMO), 2011. Updated version available at https ://sparontologies.github.io/amo/current/amo.html.

[10] Iyad Rahwan, Fouad Zablith, and Chris Reed. Laying the foundations for a World Wide Argument Web. *Artificial Intelligence*, 2007.

[11] Henri Savigny and Bruno Yun. AMELIA : A family of multi-task end-to-end language models for argumentation. *arXiv preprint arXiv :2508.17926*, 2025.

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

[13] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rauber, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv :2310.13548*, 2023.

[14] Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2003.

[15] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs : A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.