Estimer l'empreinte carbone via LLM : vers une architecture frugale et adaptative face au paradoxe énergétique.

Natalia Kalashnikova[1*], Annia Abtout[1*]
[1]Centre de Recherche et d'Innovation de Talan

natalia.kalashnikova@talan.com, annia.abtout@talan.com

Résumé

Cet article d'intention propose d'automatiser le calcul de l'empreinte carbone des appels d'offres. Si l'usage des LLM s'impose pour extraire l'information de corpus hétérogènes, il soulève un paradoxe : le coût énergétique de leur utilisation contredit le but de sobriété du projet. Ainsi, nous proposons une architecture multi-agents orchestrée qui alloue dynamiquement des ressources de calcul en fonction de la complexité informationnelle. De plus, cette étude vise à définir une métrique composite qui optimise la performance d'extraction et l'empreinte carbone.

Mots-clés

IA frugale, estimation d'empreinte carbone, IA agentique, Extraction d'informations

Abstract

This intention paper proposes to automate the carbon footprint calculation of RFPs. While the use of LLMs is necessary to extract information from heterogeneous corpora, it raises a paradox : the energy cost of their use contradicts the frugality goal of the project. To address this, we propose an orchestrated multi-agent architecture that dynamically allocates computational resources based on informational complexity. Furthermore, this study aims to define a composite metric that jointly optimizes extraction performance and carbon footprint.

Keywords

Frugal AI, carbon footprint estimation, agentic AI, information extraction

1 Introduction

L'intégration des Large Language Models (LLM) dans les processus industriels et décisionnels s'inscrit dans un paradoxe : si l'intelligence artificielle est un vecteur de performance environnementale, elle est également une source croissante de consommation de ressources (énergie, eau, métaux) [18]. Ce paradoxe se cristallise lorsque l'on cherche à mobiliser ces modèles au service du calcul d'empreinte carbone. En effet, l'outil d'optimisation environnementale devient lui-même une source de consommation qu'il convient de maîtriser. Dans ce contexte, le projet Harmonizing Optimization and Precision in Emissions (HOPE) propose une estimation *ex ante* de l'empreinte carbone des missions IT, dès la phase d'avant-vente, à partir des appels d'offres (AO). Il s'adresse en priorité aux équipes commerciales et avant-vente, qui peuvent ainsi fournir une estimation carbone dans leur réponse à AO sans mobiliser d'expertise environnementale dédiée. Un appel d'offres est un dossier de documents qui explique le besoin d'un acheteur (souvent une entreprise) afin de recevoir des solutions commerciales et techniques (par exemple, l'achat de biens, la réalisation de travaux, la prestation de services, etc.). Notre dataset est constitué des appels d'offres que Talan a reçus dans le cadre de son activité. Cette estimation précoce s'inscrit dans un contexte où les entreprises sont soumises à des obligations croissantes de transparence environnementale, portées notamment par la directive européenne CSRD (Corporate Sustainability Reporting Directive)[1]. Dans ce cadre, l'empreinte carbone s'impose progressivement comme critère d'évaluation dans les AO du secteur IT, rendant son estimation dès la phase d'avant-vente stratégique. Rendre visible cet impact dès la conception des offres constitue par ailleurs un levier de sensibilisation pour les équipes projet. Les AO présentent cependant des défis spécifiques : documents longs, hétérogènes, sans format unifié, dans lesquels l'information nécessaire au calcul de l'empreinte est fréquemment implicite, incomplète ou absente. Pour structurer cette estimation, HOPE s'appuie sur un formulaire carbone dont les champs correspondent aux principaux postes d'émission d'une mission IT (mobilité, hébergement, infrastructure de calcul, achats). Or, renseigner ce formulaire manuellement à partir des AO est chronophage et source d'erreurs : c'est précisément ce que l'automatisation doit corriger, sans introduire de nouvelles inexactitudes. Le recours aux LLM s'impose pour traiter ces corpus non structurés, mais soulève précisément le paradoxe initial. La question n'est donc pas de savoir s'il faut utiliser l'IA, mais *comment* le faire sans que le remède ne devienne le problème. C'est le positionnement central de l'IA frugale [6] : non pas rejeter l'usage de l'IA mais n'y recourir que lorsque cela est strictement nécessaire à la performance, en questionnant systématiquement chaque choix architectural au regard de son coût énergétique. Cet

*. Ces auteurs ont contribué de manière égale à ce travail.

1. https ://eur-lex.europa.eu/legal-content/EN/TXT/ ?uri=CELEX :32022L2464

article propose une architecture expérimentale pour explorer ce compromis, accompagnée d'une métrique composite conçue pour évaluer conjointement la qualité d'extraction et la sobriété énergétique.

2 État de l'art

2.1 Extraction d'information à partir de documents non structurés

Plusieurs stratégies visent à optimiser la gestion du contexte. LoongServe [20] exploite *elastic sequence parallelism* qui permet d'ajuster dynamiquement le nombre de GPU alloués à une requête, ce qui permet de traiter des contextes plus longs, tandis qu'InfiniPot [12] utilise une gestion locale de la mémoire pour fonctionner sur des ressources limitées. Ces techniques améliorent la gestion du contexte des documents longs et bien structurés, mais leur efficacité diminue fortement dans des contextes plus désorganisés, ce qui est typique des AO. En parallèle, l'approche de [16] propose de contourner le découpage explicite (chunking) en permettant une récupération directe dans le texte, ce qui évite les erreurs de segmentation. Toutefois, cette approche demeure instable pour les documents denses. Enfin, le chunking sémantique présente un surcoût computationnel difficilement justifiable [17]. Les approches analysées sont encore insuffisamment robustes pour des déploiements en production dans des cas d'usage complexes comme les AO. Dans ce contexte, on s'est orienté vers l'approche de l'extraction d'information par prompting, comme PromptNER [1] ou l'enrichissement des prompts avec des connaissances externes [2]. Ces méthodes permettent d'extraire des entités sans phase d'apprentissage. En plus, elles sont simples à déployer et flexibles, mais fortement dépendantes de la formulation du prompt. Les entités complexes ou ambigües sont mal reconnues en l'absence d'exemples concrets. Pour le projet HOPE, les approches par prompt offrent un bon compromis entre simplicité de mise en œuvre et performance. Leur capacité à être rapidement adaptées à différents cas d'usage en fait une piste d'exploration prioritaire, même si leur robustesse reste à renforcer.

2.2 Raisonnement déductif via LLM

Le cadre de l'approche de prompting permet d'explorer des stratégies de raisonnement guidé : prompting en chaîne de pensée [3], raisonnement inductif et déductif [4] et structuration syllogistique [19] afin de déduire des informations implicites à partir des AO.

Ces méthodes montrent une efficacité pour des raisonnements bien encadrés, notamment lorsque le prompt fournit une structure logique explicite. Néanmoins, dès que les éléments nécessaires au raisonnement sont implicites, l'efficacité des modèles diminue. Le recours à des logiques formelles ou des chaînes déductives pourrait enrichir les prompts dans HOPE, mais nécessite encore une exploration approfondie pour garantir la cohérence et la généralisation des résultats.

2.3 Enjeux de l'IA frugale et son intégration avec les LLM

L'estimation du coût environnemental lié à l'utilisation des LLM représente plusieurs difficultés. Tout d'abord, elle dépend de la localisation des centres de données où le modèle utilisé est déployé par rapport au mode de production de l'électricité. De plus, la plupart des fournisseurs des modèles les plus utilisés ne divulguent ni l'empreinte carbone liée à l'entraînement et à l'inférence des modèles, ni le nombre de requêtes par jour.

Toutefois, la littérature scientifique s'enrichit progressivement d'études visant à quantifier l'empreinte carbone des LLM. Ainsi, [21] et [11] estiment qu'une requête consomme environ 0,34 Wh. À son tour, Mistral propose des initiatives pour unifier l'estimation de l'impact environnemental des modèles. Selon leur étude, l'entraînement et 18 mois d'utilisation du modèle Mistral Large 2 ont généré 20,4 ktCO2e et 281 000 m3 d'eau consommée, soit l'équivalent de 4000 aller/retour Paris-Tokyo[2]. Leur analyse estime que la génération d'une page par ce modèle émet 1,14 g de CO2e (soit 10 secondes de streaming d'une vidéo aux États-Unis ou 55 secondes en France) et consomme 0,05 L d'eau (soit l'eau nécessaire pour pousser un radis rose) [15].

L'intégration des LLM dans différents processus de l'entreprise permet d'automatiser des tâches manuelles et chronophages mais il faut tenir compte de la maîtrise du coût environnemental de leur utilisation. Dans ce contexte, la recherche s'oriente vers la "frugalité numérique", cherchant à trouver le compromis entre la performance des modèles et l'empreinte carbone de leur utilisation.

La littérature identifie trois leviers principaux pour atteindre cette frugalité numérique. Le premier concerne le dimensionnement optimal des modèles. Les études montrent que les modèles légers et locaux suffisent pour la plupart des tâches professionnelles [10]. Pour les cas d'usage particuliers, les modèles spécialisés pour une tâche donnée consomment moins d'énergie tout en offrant une performance comparable aux LLM géants [14].

Le deuxième levier repose sur l'optimisation logicielle et matérielle lors de l'inférence. Dans le cas spécifique des systèmes de génération augmentée par récupération (RAG), l'ajustement des seuils de similarité et la réduction de la dimension des embeddings sont identifiés comme des stratégies clés pour diminuer la charge de calcul [9].

Enfin, la littérature souligne la nécessité de cadres de mesure standardisés et multicritères. Ainsi, des modèles récents comme LLMCarbon intègrent l'empreinte "incarnée" (fabrication du matériel) et les spécificités des architectures modernes (MoE) pour offrir une vision globale du bilan carbone [7].

HOPE n'est pas le premier outil qui utilise les LLM pour les enjeux environnementaux. Par exemple, Wang et al. [5] ont proposé l'approche combinée de LLM et RAG qui permet une mise à jour dynamique des données, notamment

2. Calculé à l'aide du simulateur de l'ADEME : https ://impactco2.fr/outils/comparateur

pour l'Analyse du Cycle de Vie. La méthode surpasse les approches traditionnelles et d'autres configurations LLM, fournit des taux de récupération d'information plus élevés et des écarts d'estimation plus faibles, l'efficacité économique est accrue car le système évite des mises à jour fréquentes des paramètres du LLM.

3 Problématique de recherche

L'analyse de l'état de l'art met en évidence une tension entre la nécessité d'avoir recours à des modèles performants capables de raisonnement, souvent coûteux en énergie, pour traiter des documents hétérogènes tels que les AO, et notre objectif d'estimation carbone, qui impose une contrainte forte de frugalité. La problématique au cœur de ce projet est donc de concilier la performance d'extraction nécessaire à la fiabilité du bilan carbone avec la sobriété énergétique de l'inférence. Les solutions actuelles tendent à polariser la solution avec, d'un côté, des modèles légers mais limités aux tâches d'extraction complexes, et de l'autre, des LLM massifs qui surperforment le problème. Or, la réalité d'un AO est plus nuancée, car ce type de document combine des données simples (dates, lieux) et des exigences techniques complexes. Pour répondre à ce double impératif, nous formulons l'hypothèse d'une architecture hybride, capable de solliciter les modèles de haute performance de manière conditionnelle et sélective, permettrait d'optimiser le ratio performance/empreinte carbone.

4 Méthodologie

Pour valider cette hypothèse, nous proposons HOPE, une architecture modulaire dont le principe central est d'allouer dynamiquement les ressources computationnelles en fonction de la difficulté d'extraction de chaque champ, en s'appuyant sur des connaissances métier structurées.

La Figure 1 illustre l'architecture que nous proposons pour répondre à notre problématique. Dans un premier temps, afin de prendre en compte la complexité des AO composés de plusieurs types d'éléments différents et de récupérer le maximum d'informations, nous extrayons séparément le texte et les images. Les images sont envoyées à un LLM multimodal afin d'être transcrites et cette transcription est ensuite insérée à la place de l'image. Le contenu extrait de chaque page est envoyé à un modèle d'embeddings. Les embeddings de chaque page sont ensuite indexés dans une base vectorielle pour pouvoir faire de la recherche par similarité sémantique. Toujours dans un but de frugalité, nous allons comparer plusiers modèles d'embeddings afin d'identifier celui présentant le meilleur rapport entre performance et impact environnemental. Cette étape constitue le module de pré-traitement des AO.

Dans un deuxième temps, pour pallier les difficultés des Small Language Models (SLM) à gérer l'implicite sans contexte, nous introduisons un module de métadonnées expert (ECO-Metadata), inspiré des travaux de [22]. Ce module optimise la sélection des passages pertinents du document avant l'étape d'extraction et fournit au modèle un contexte sémantique spécialisé. Pour chaque champ du formulaire carbone, nous définissons un quintuplet `<Thématique, Sujet, Métrique, Mots-clés, Connaissances>`. Les quatre premiers éléments guident l'extraction des informations explicites, tandis que le champ *Connaissances* permet de déduire des informations implicites à partir de règles métier contextuelles, sans recourir systématiquement à un LLM coûteux. Afin d'illustrer concrètement le fonctionnement de l'ECO-metadata et de la cascade d'extraction, nous détaillons le traitement du champ *Infrastructure de calcul* (cf. prompt 1). En effet, ce champ présente une complexité informationnelle caractéristique des AO où l'information est rarement explicite et nécessite souvent une déduction à partir d'indices contextuels.

Listing 1 – Prompt Niveau 2

```
Tu es un assistant specialise en extraction d'informations
pour les bilans carbone de projets informatiques.

## Contexte metier
Thematique : Numerique & Infrastructure de calcul
Sujet      : Type de ressources computationnelles mobilisees
Metrique   : Type d'infrastructure (GPU/CPU/Cloud/On-premise),
             fournisseur si cloud, localisation si mentionnee
Mots-cles  : GPU, CPU, cloud, serveur, infrastructure,
             AWS, Azure, GCP, on-premise, datacenter,
             entrainement, inference

## Extrait de document

## Instruction
Extrais uniquement les informations relatives au type
d'infrastructure de calcul explicitement mentionnees.
Ne deduis pas ce qui n'est pas ecrit.

Reponds uniquement en JSON :
{
  "type_infrastructure" : "GPU|CPU|Cloud|On-premise|null",
  "fournisseur_cloud"   : "string|null",
  "localisation"        : "string|null",
  "source_textuelle"    : "passage exact justifiant l'extraction"
}
```

Le cœur de notre contribution réside dans le Routeur de Complexité appliqué dans le module de l'Extraction de l'information. Plutôt que d'appliquer une inférence coûteuse uniformément, le système évalue la difficulté d'extraction de chaque champ et dirige la requête vers la méthode la plus adaptée (cf. Figure 1, bloc central) :

- **Niveau 1 (Symbolique) :** Pour les données explicites et standardisées (dates, lieux), une extraction par expressions régulières ou recherche de mots-clés est priorisée. Le coût énergétique est négligeable.
- **Niveau 2 (Inférence Frugale - SLM) :** Pour l'extraction sémantique, nous mobilisons des Small Language Models exécutables localement. Le choix du modèle sera guidé par un benchmark évaluant le compromis performance/empreinte carbone. Le module ECO-metadata enrichit le contexte de ces modèles afin d'améliorer la qualité d'extraction sans compromettre l'objectif de frugalité.
- **Niveau 3 (Raisonnement Complexe - LLM) :** Uniquement en cas d'échec des niveaux précédents ou de détection d'ambiguïté forte, la requête est transmise à un LLM performant. À ce niveau, nous envisageons d'évaluer plusieurs stratégies de raisonnement issues

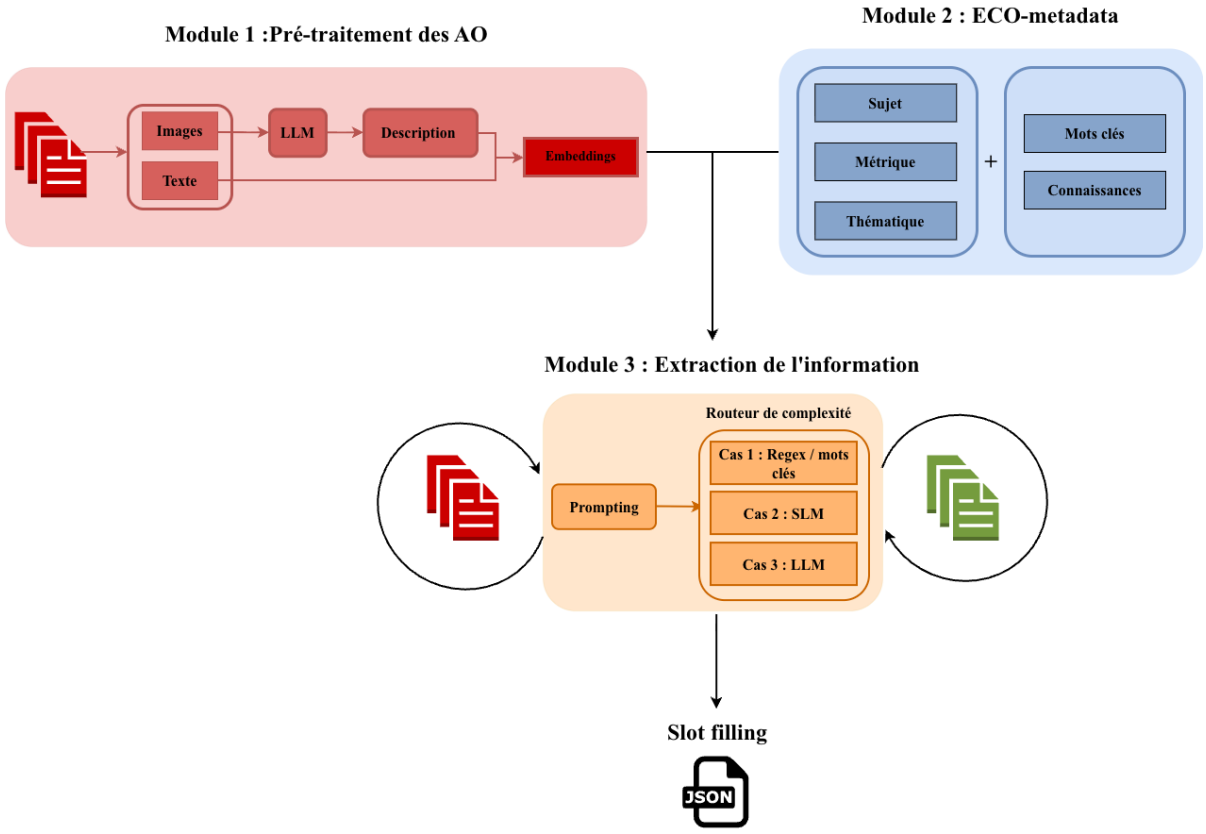

FIGURE 1 – Architecture modulaire du système HOPE. Le flux de données s'articule autour de trois modules principaux : (1) le prétraitement des AO, (2) le module ECO-metadata pour l'enrichissement sémantique des requêtes par l'expertise métier, et (3) le moteur d'extraction d'informations par RAG. Ce dernier traite le document d'AO et la documentation experte pour générer, en sortie, un fichier JSON des informations extraites de l'AO destiné à l'estimation de son empreinte carbone.

de la littérature, telles que CoT ou la structuration syllogistique, afin d'identifier celle offrant le meilleur compromis entre capacité d'inférence implicite et coût énergétique.

La transition entre niveaux repose sur un score de confiance : un seuil statique, fixé expérimentalement, gouverne le passage du Niveau 1 au Niveau 2 ; la transition vers le Niveau 3 est quant à elle déléguée à un agent orchestrateur de type *LLM-as-a-judge*, chargé d'évaluer la pertinence des réponses du SLM et de déclencher, si nécessaire, une extraction plus coûteuse sur les champs complexes ou implicites. Ce composant mobilise lui-même un LLM, ce qui introduit un coût additionnel que notre métrique d'éco-précision permettra de quantifier. Nous faisons l'hypothèse que ce coût reste inférieur à celui d'une escalade systématique vers le Niveau 3, en limitant les appels à ce dernier aux seuls champs pour lesquels le SLM a effectivement échoué. Au vu de la nature des AO, leur analyse présente souvent des lacunes informationnelles, avec des champs peu ou pas renseignés. Afin de pallier à ce cas de figure courant nous proposons d'appliquer la procédure d'extraction d'information via le routeur de complexité sur une base documentaire distincte composée de documentation experte telle que la Base Carbone ADEME, les normes ISO ou des référentiels internes. Le JSON obtenu sera alors évalué face à la vérité terrain, dans une démarche d'optimisation expérimentale de notre cas d'application. L'évaluation de notre système reposera sur une métrique composite pondérant le F1-Score classique par l'empreinte carbone de l'inférence. Dans un premier temps, nous limitons les entrées traitées aux documents au format pdf. Cependant, HOPE prendra en charge d'autres formats à terme.

5 Protocole d'Évaluation

Afin d'évaluer la pertinence de notre architecture, nous conduisons une campagne de benchmarking sur un corpus de 100 AO. Chaque domaine de l'industrie, voire chaque entreprise dispose de son propre modèle pour constituer le dossier des AO, ce qui rend ces documents structurellement très hétérogènes. La constitution du corpus a donc été guidée par deux critères complémentaires. D'une part, la diversité des domaines, afin de couvrir des formats variés et de refléter l'activité réelle de l'entreprise. D'autre part, la variété des tailles de documents, afin d'évaluer la sensibilité des modèles à la longueur du contexte et la robustesse des stratégies de chunking. Ainsi, le jeu de données diversifié permettra d'éprouver la robustesse de notre système ainsi que la généralisation de l'architecture au-delà d'un

cas d'usage unique. Nous avons donc développé une plateforme d'annotation afin d'étiqueter les AO et de construire le référentiel d'évaluation. Le processus consiste à parcourir chaque document et à renseigner un formulaire structuré dont les champs correspondent à ceux du formulaire carbone final. Les annotateurs peuvent indiquer si l'information est absente du document et s'ils l'ont déduite à partir d'autres informations présentes dans le document ou de leurs propres connaissances. Pour chaque champ renseigné, le niveau de fiabilité de l'information saisie est indiqué. Chaque document est annoté par trois personnes, et la version finale du document annoté est constituée du vote majoritaire des trois annotations. L'objectif est de comparer le JSON qui agrège les champs pertinents généré par HOPE avec ce référentiel expert. Nous proposons une approche hybride afin d'évaluer notre architecture, conçue pour être cohérente avec le principe de frugalité qui structure l'ensemble de notre projet. Nous distinguons donc deux régimes d'évaluation selon la nature des champs du formulaire carbone. L'objectif de cette approche hybride est de limiter l'utilisation de méthodes d'évaluation coûteuses aux champs le nécessitant uniquement.

Ainsi, pour les champs à valeur déterministe et explicite (tels que les dates, lieux, etc.) la comparaison entre le JSON produit par HOPE et la vérité terrain est réalisée via des métriques classiques d'extraction d'information : la Précision, le Rappel, le F1-score et l'Exact Match (EM) au niveau des valeurs extraites. Ces métriques sont bien adaptées à la nature binaire de ces champs et leur coût computationnel est très faible. Concernant les champs nécessitant une inférence sémantique, nous adoptons le framework d'évaluation proposé par [8]. Ce protocole combine une évaluation quantitative par métriques (BERTScore, ROUGE, F1) et une évaluation qualitative par LLM-juge. Un mécanisme de détection de conflits compare les deux scores et si leur écart dépasse un seuil, une validation humaine experte est déclenchée. Bien que plus énergivore, ce protocole est adapté aux champs pour lesquels la réponse est implicite dans le document et permet de garantir une évaluation robuste de notre architecture avant sa mise en production. De plus, cette approche a été proposée dans le cadre de l'évaluation des documents de l'entreprise, ce qui correspond à notre contexte de travail.

Conformément à l'objectif de notre projet, afin de quantifier l'impact d'une architecture plus frugale sur l'extraction de données depuis des documents hétérogènes, nous définissons une métrique composite, le score d'*éco-précision*. Pour chaque champ i du formulaire carbone, nous avons un score d'extraction s_i. Le score hybride global $Score_{hybride}$ correspond à l'agrégation de ces scores sur l'ensemble des champs. L'éco-précision est alors formulée ainsi :

$$\text{Éco-précision} = \frac{score_{hybride}}{\log(1 + empreinte_{carbone})} \quad (1)$$

Où $Score_{hybride}$ correspond à la mesure de l'exactitude de l'extraction. L'empreinte carbone de l'inférence est calculée à l'aide de sondes logicielles telles que CodeCarbon [13].

Cette métrique pénalise les architectures qui, bien que performantes selon le framework de validation croisée, nécessiteraient une consommation énergétique disproportionnée. Aussi, la consommation énergétique des trois niveaux du routeur s'étendant sur plusieurs ordres de grandeur, une compression logarithmique est nécessaire pour préserver la comparabilité entre configurations. Ce choix sera éprouvé expérimentalement par une analyse de sensibilité sur le corpus annoté.

6 Conclusion

Cet article s'inscrit dans une démarche de questionnement systématique de l'usage de l'IA : plutôt que de recourir uniformément aux modèles les plus performants, nous défendons l'idée que chaque étape architecturale doit être conçue pour mobiliser le niveau de ressources strictement nécessaire à la tâche. C'est ce principe qui structure HOPE, une architecture modulaire frugale pour l'extraction automatique des informations nécessaires à l'estimation *ex ante* de l'empreinte carbone des missions IT à partir des AO. Face au paradoxe énergétique que soulève le recours aux LLM dans un contexte de sobriété, nous proposons un routeur de complexité à trois niveaux qui alloue dynamiquement les ressources computationnelles en fonction de la difficulté informationnelle de chaque champ, enrichi par un module ECO-metadata qui ancre l'extraction dans une expertise métier structurée. L'évaluation de cette architecture repose sur une métrique composite, l'éco-précision, qui évalue conjointement la qualité d'extraction et l'empreinte carbone de l'inférence. Cette proposition appelle plusieurs axes de validation expérimentale. La calibration empirique des seuils de confiance du routeur sur le corpus annoté de 100 AO constituera une première étape clé, de même que la comparaison systématique des configurations de modèles selon la métrique éco-précision. Par ailleurs, si l'architecture a été conçue dans le contexte spécifique des AO du secteur IT, sa logique de routage adaptatif pourrait se généraliser à d'autres corpus documentaires hétérogènes pour lesquels la tension entre performance d'extraction et sobriété énergétique se pose de manière similaire. Par ailleurs une limite inhérente à ce stade de la recherche est que l'architecture proposée n'a pas encore été validée expérimentalement et nécessite une calibration rigoureuse sur données réelles. Une fois la configuration optimale identifiée, le JSON des champs extraits a vocation à alimenter un calculateur dédié qui produira l'estimation finale de l'empreinte carbone de la mission, à partir de règles de calcul fondées sur les référentiels ADEME. Cela constituant ainsi le déploiement opérationnel visé par le projet HOPE. Ces travaux feront l'objet de publications ultérieures.

Remerciements

Merci à Sophie Fayad et Steve Bellart, qui ont participé à l'instigation du projet. Ainsi qu'à André Darmedru, Imane Aregabi, Arnaud Metras, Elyes Fakhar et Djida Oubekkou, dont l'expertise a contribué au bon avancement du projet. Enfin, merci à Laurent Cervoni, directeur du centre RetI

de Talan, pour son soutien constant et son engagement en faveur de la recherche.

Références

[1] D. Ashok and Z. C. Lipton. Promptner : Prompting for named entity recognition, 2023.

[2] J. Bian, J. Zheng, Y. Zhang, H. Zhou, and S. Zhu. One-shot biomedical named entity recognition via knowledge-inspired large language model. BCB '24, New York, NY, USA, 2024. Association for Computing Machinery.

[3] Z. Chu et al. Navigate through enigmatic labyrinth a survey of chain of thought reasoning : Advances, frontiers and future. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1173–1203, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[4] C. Cai et al. The role of deductive and inductive reasoning in large language models, 2025.

[5] H. Wang et al. Carbon footprint accounting driven by large language models and retrieval-augmented generation, 2024.

[6] M. Evchenko. Frugal learning : Applying machine learning with minimal resources.

[7] A. Faiz, S. Kaneda, R. Wang, R. Osi, P. Sharma, F. Chen, and L. Jiang. Llmcarbon : Modeling the end-to-end carbon footprint of large language models, 2024.

[8] F. Grina and N. Kalashnikova. Évaluation de la robustesse des llm : Proposition d'un cadre méthodologique et développement d'un benchmark. In *Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, pages 151–163, 2025.

[9] Z. Guo, C. Gao, and J. Bogner. On the effectiveness of proposed techniques to reduce energy consumption in rag systems : A controlled experiment, 01 2026.

[10] J. Haase, F. Klessascheck, J. Mendling, and S. Pokutta. Sustainability via llm right-sizing, 2025.

[11] N. Jegham, M. Abdelatti, C. Young Koh, L. Elmoubarki, and A. Hendawi. How hungry is ai ? benchmarking energy, water, and carbon footprint of llm inference, 2025.

[12] M. Kim, K. Shim, J. Choi, and S. Chang. InfiniPot : Infinite context processing on memory-constrained LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16046–16060, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[13] K. Lottick, S. Susai, S. A. Friedler, and J. P. Wilson. Energy usage reports : Environmental awareness as part of algorithmic accountability. *arXiv preprint arXiv :1911.08354*, 2019.

[14] S. Luccioni, Y. Jernite, and E. Strubell. Power hungry processing : Watts driving the cost of ai deployment ? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 85–99, New York, NY, USA, 2024. Association for Computing Machinery.

[15] Mistral AI. Our contribution to a global environmental standard for ai, July 2025. Consulté le 5 mars 2026.

[16] H. Qian et al. Grounding language model with chunking-free in-context retrieval. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1298–1311, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[17] Renyi Qu, Ruixuan Tu, and Forrest Sheng Bao. Is semantic chunking worth the computational cost ? In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics : NAACL 2025*, pages 2155–2177, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[18] A. Singh, N. P. Patel, A. Ehtesham, S. Kumar, and T. T. Khoei. A survey of sustainability in large language models : Applications, economics, and challenges. In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 00008–00014. IEEE, 2025.

[19] W. Wan et al. Sr-fot : a syllogistic-reasoning framework of thought for large language models tackling knowledge-based reasoning tasks. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025.

[20] B. Wu et al. Loongserve : Efficiently serving long-context large language models with elastic sequence parallelism. In *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles*, SOSP '24, page 640–654, New York, NY, USA, 2024. Association for Computing Machinery.

[21] J. You. How much energy does chatgpt use ?, February 2025. Consulté le 5 mars 2026.

[22] Y. Zou et al. Esgreveal : An llm-based approach for extracting structured data from esg reports. *Journal of Cleaner Production*, 489 :144572, 2025.