Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Transfert de modèles de langue pour la classification rhétorique des citations à travers les disciplines

Anne-Sophie Foussat[123] Vincent Guigue[2] Nicolas Sauvion[3] Robert Bossy[1]
Claire Nédellec[1]
(1) Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France
(2) AgroParisTech, MIA Paris-Saclay, Palaiseau, France
(3) PHIM, Univ Montpellier, INRAE, CIRAD, IRD, Institut Agro, Montpellier, France
anne-sophie.foussat@inrae.fr, robert.bossy@inrae.fr,
vincent.guigue@agroparistech.fr, nicolas.sauvion@inrae.fr,
claire.nedellec@inrae.fr

RÉSUMÉ

La classification automatique des fonctions rhétoriques des citations contribue à l'étude des stratégies discursives d'un auteur lorsqu'il cite, et plus généralement, de son intention. Dans l'objectif d'estimer la fiabilité des découvertes citées en écologie, cet article analyse les capacités de transfert des modèles de langue affinés en linguistique computationnelle pour cette tâche, en les comparant aux méthodes par amorçage (*prompting*). Nous introduisons PD100cit, un nouveau corpus annoté, ainsi qu'un guide d'annotation, afin d'explorer la typologie rhétorique des citations relatives aux interactions biologiques. Nous explorons également la sensibilité des modèles aux longueurs des contextes des passages de citations. Nos résultats montrent de bonnes performances des modèles de langue transférés en écologie et l'intérêt de réviser la typologie pour évaluer la fiabilité des découvertes de la linguistique computationnelle à l'écologie.

ABSTRACT

Transfer of language models for rhetorical citation classification across disciplines.

The automatic classification of rhetorical function examines an author's discursive strategy when citing and, more broadly, their intent. This article investigates the transferability of language models, fine-tuned in computational linguistics. Fine-tuned language models are compared to prompting-based approaches to assess the reliability of cited findings in ecology. We introduce PD100cit, a newly annotated dataset, alongside annotation guidelines, to explore the rhetorical typology of citations related to biological interactions. We also explore the models' sensitivity to citation context sizes. Our results show strong performances of language models in ecology, though further refinements in the typology are necessary to evaluate the reliability of cited findings from computational linguistics to ecology.

MOTS-CLÉS : classification de citations, modèle de langue, interactions biologiques.

KEYWORDS: citation classification, language models, biologic interactions.

ARTICLE : **Soumis à** CORIA-TALN 2025 /CORIA.

1 Introduction

Dans un contexte de profusion de publications scientifiques, nous nous intéressons à la fiabilité de nouvelles découvertes à travers leur perception par la communauté scientifique, telle qu'elle s'exprime dans les citations. La construction des connaissances scientifiques repose sur des dynamiques par lesquelles les découvertes émergent, sont diffusées, discutées, puis progressivement acceptées ou rejetées. Analyser l'évolution de la reconnaissance et de la controverse autour d'une découverte permet d'évaluer son acceptation par la communauté scientifique, et de mieux comprendre si elle est perçue comme fiable et consensuelle. Chaque reprise de cette découverte, par la citation, constitue une trace de son parcours dans la littérature scientifique et un indice de son appropriation ou de sa remise en question. A travers la citation, un auteur peut réfuter, confirmer ou reconnaître une découverte, phénomène qui permet d'étudier l'état des lieux de la connaissance et des consensus autour des découvertes scientifiques. Par exemple, cette citation par Jarausch et al. (2019) montre que leur résultat confirme des travaux antérieurs - la plante hôte préférée d'un insecte [*Cacopsylla pruni*] est le prunellier [*Prunus spinosa*]) :

> "... it is the preferred host plant of *C. pruni* (Lauterer 1999) and accordingly, we found *C. pruni* on every *P. spinosa* tested sometimes at high population densities. This is supported by data from Carraro et al. (2002), Yvon et al. (2004) or Maier et al. (2013)".

Établir la vraisemblance d'une découverte en science à travers les citations est une tâche ardue, reposant en partie sur l'interprétation individuelle des auteurs d'un socle de connaissances plus ou moins consolidées. En 1986, Swales a souligné que la construction d'un texte académique est marquée par la subjectivité de l'auteur, car sa propre expertise dans son domaine influence son intention, son point de vue et ses choix de langage (Swales, 1986). Derrière le discours se cachent ainsi des intentions rhétoriques pour informer et convaincre le lecteur par une analyse critique des travaux antérieurs. Swales a proposé un cadre pour classer ces intentions et saisir la part de subjectivité.
Un autre défi pour objectiver une découverte est la variabilité des domaines scientifiques. Ainsi, Hyland (1999) a souligné que la validation des connaissances et l'utilisation des citations varient selon les disciplines, influençant la structure du texte et les stratégies d'argumentation. De ce fait, la citation devient un mécanisme reflétant les normes disciplinaires. La classification automatique des citations par apprentissage doit donc s'appuyer sur des modèles robustes, capables de s'adapter à ces spécificités.
Différentes classifications ont été proposées pour catégoriser automatiquement la fonction rhétorique des citations (Teufel et al., 2006 ; Zhang et al., 2022 ; Jiang & Chen, 2023), mais la transférabilité des classifieurs et des classes entre domaines reste à approfondir. Si les stratégies discursives varient selon les disciplines, l'impact de ces variations sur le transfert des modèles de classification de citations et de typologie des citations demeure peu étudié. Or, dans les domaines spécialisés où les données sont limitées, ce transfert est crucial pour évaluer la fiabilité des découvertes citées.
Cet article analyse la capacité de transfert des modèles SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), RoBERTa (Liu et al., 2019) et BioLinkBERT (Yasunaga et al., 2022) dans deux disciplines distinctes : la linguistique computationnelle et l'écologie, science qui étudie les relations entre les organismes vivants et leur environnement. Notre cas d'étude est une maladie des arbres fruitiers, le dépérissement du poirier, ou *Pear decline*, liée à des interactions biologiques complexes. Ce sujet se caractérise par un nombre de publications conséquent, une forte évolutivité des connaissances, parfois incertaines voire contradictoires, et une communauté scientifique restreinte. Nous introduisons le corpus PD100cit centré sur ce sujet, qui est disponible sur GitHub [1] avec le guide d'annotation et

1. https://github.com/AnneSophie148/PhD/tree/main/citation_classification

les modèles de classification rhétorique des citations.
Nous allons montrer que les modèles de langue affinés pour la classification rhétorique des citations peuvent être appliqués à l'étude des interactions biologiques en répondant aux questions de recherche suivantes : la typologie des fonctions rhétoriques des citations en linguistique computationnelle est-elle adéquate et suffisante pour l'étude des interactions biologiques ? Les modèles développés en linguistique computationnelle peuvent-ils être transposés à l'étude des citations de travaux des interactions biologiques ?

2 Etat de l'art

Moravcsik et Murugesan (1975) ont établi une première typologie des citations, distinguant des autres citations celles des découvertes considérées comme une avancée significative pour la science. Ils proposent une classification selon quatre axes : (i) conceptuel ou opérationnel (concept théorique ou usage) ; (ii) organique ou superficiel (citation essentielle ou seule mention de l'existence d'un travail) ; (iii) évolutif ou juxtapositionnel (l'article qui cite est un prolongement des travaux cités ou une alternative) ; (iv) confirmatif ou négatif (confirmation ou critique des travaux cités). Cette classification nuance la reprise d'information et distingue les citations confirmant une découverte de celles qui les remettent en question. Ces travaux ont inspiré diverses caractérisations des citations par leurs fonctions rhétoriques (Teufel et al., 2009), leur importance (Valenzuela et al., 2015 ; Wan & Liu, 2014), leur contexte (Hassan et al., 2018), ou encore leur polarité (Abu-Jbara et al., 2013). Cet article se concentre sur la classification des citations par leurs fonctions rhétoriques en linguistique computationnelle et en écologie en tant qu'indice discursif permettant d'analyser la manière dont une découverte est reçue, discutée ou contestée dans le temps, et permettant ainsi d'évaluer sa fiabilité au sein de la communauté scientifique.

2.1 La classification de la fonction rhétorique

La classification de la fonction rhétorique consiste à classer la citation par son rôle dans la structure de l'argumentation d'un article. Cette tâche a fait l'objet de nombreuses études, évoluant des approches à base de règles (Garzone & Mercer, 2000), vers des approches par apprentissage (Teufel et al., 2006). Teufel et al. (2006) se basent sur des caractéristiques linguistiques, telles que les parties du discours, le temps et la voix des verbes, pour proposer une classification à 12 classes. Plus récemment, Jurgens et al. (2018) approfondissent cette classification en combinant caractéristiques linguistiques, bibliométriques et structurelles, et proposent le jeu de données ACL-ARC. Cohan et al. (2019) présentent Sci-Cite, un autre jeu de données pour une classification à seulement trois classes. Avec l'essor de l'apprentissage profond, Zhang et al. (2022) introduisent le corpus NI-Cite pour une classification à trois classes avec le modèle SciBERT. Pour cela, ils combinent les représentations du passage de citation à des métadonnées du document. Jiang et Chen (2023) explorent également les performances de SciBERT en proposant le corpus Jiang2021. Leur classification de 11 classes est adaptée du schéma de 12 classes de Teufel et al. (2009). Jiang2021 fusionne les annotations de fonctions rhétoriques de plusieurs corpus de linguistique computationnelle. Dans le domaine biomédical, d'autres approches analysent la fonction des citations (Agarwal et al., 2010). Toutefois les classifications varient d'une étude à une autre, et les jeux de données ne sont pas accessibles (Jiang & Chen, 2023).

2.2 Spécificité des domaines

Les stratégies argumentatives des articles scientifiques varient selon les disciplines scientifiques (Harwood, 2009 ; Hyland, 1999) avec une distinction entre les domaines des sciences exactes et les autres. Par exemple, Hu et Wang (2014) analysent l'influence disciplinaire sur les pratiques de citation dans les articles de recherche, en comparant la linguistique appliquée et la médecine. Leur étude révèle une prédominance des citations non intégrées en médecine, où l'auteur cité n'apparaît pas directement dans la structure syntaxique de la phrase, ainsi qu'un usage fréquent de verbes factuels. En revanche, en linguistique appliquée, les citations intégrées et des verbes plus subjectifs sont privilégiés, laissant place au débat scientifique et à l'interprétation. D'autres disciplines sont comparées par Zheng et Li (2022), dont la linguistique, la biologie, la physique et l'éducation. ll observe que la biologie se distingue par une forte densité de citations dans les sections Résultats et Discussion, et moins de diversité des formes de citation. En analyse des citations interdisciplinaires en sciences dures, Bornmann (2019) examine les concepts cités par les articles scientifiques faisant référence aux œuvres de Popper. L'étude montre que le concept de corroboration de Popper, selon lequel une hypothèse scientifique est dite corroborée tant qu'elle résiste aux tests susceptibles de la falsifier, est plus fréquemment présent en biologie, discipline empirique, que dans d'autres disciplines, y compris l'informatique. Cet article reprend la classification des fonctions rhétoriques telle que proposée par Jiang et Chen (2023), et explore son application dans le domaine de l'écologie. Nous examinons l'adéquation de la typologie des fonctions rhétoriques des citations à l'analyse des dynamiques de circulation et d'acceptation des connaissances scientifiques dans cette discipline, à travers l'étude du *Pear Decline*, un cas marqué par une diversité d'organismes vecteurs de pathogènes et une disponibilité limitée de données structurées.

2.3 Classification de la fonction rhétorique par *prompt*

La classification par *prompt*, soutenue par l'essor des grands modèles de langue (LLM) est une approche de plus en plus utilisée en linguistique computationnelle. Kunnath et al. (2023) l'ont appliquée à l'analyse des citations dans les articles scientifiques, comparant GPT-3 en *zéro-shot* à SciBERT affiné. GPT-3 obtient de meilleures performances sur le jeu de données multidisciplinaire ACT2 (Kunnath et al., 2022), tandis que SciBERT obtient de meilleurs résultats sur ACL-ARC (Jurgens et al., 2018). Des travaux récents explorent davantage de modèles pour la classification de citations (Koloveas et al., 2025). Parmi les modèles Llama 3, Mistral Nemo, Phi 3 Medium, Phi 3.5 Mini, Gemma 2, Qwen 2, et Qwen 2.5 utilisés, Qwen 2.5 obtient les meilleures performances pour cette tâche en *zero-shot*, *one-shot*, *few-shot*, et *many-shot*. Cet article prolonge la comparaison entre GPT et des modèles de langue affinés sur la classification des citations, en appliquant la typologie de Jiang et Chen (2023) au domaine de l'écologie.

3 Matériel et méthodes

3.1 Les corpus

Nous avons utilisé le corpus Jiang2021 en linguistique computationnelle pour entraîner les modèles de langue pour classer les fonctions rhétoriques des citations. Il contient 3 356 citations en contexte,

obtenues sur demande auprès des auteurs. Ce corpus a été choisi pour la diversité des classes rhétoriques, en particulier celles qui expriment une posture argumentative (comme les classes "support", "contrast" et "weakness"), essentielles à l'analyse de la fiabilité des connaissances scientifiques. Jiang2021 regroupe six jeux de données précédents : Teufel2010 (Teufel, 2010), Dong2011 (Dong Schäfer, 2011), Jha2016 (Abu-Jbara et al., 2013 ; Jha et al., 2016), Alvarez2017 (Hernández-Alvarez et al., 2017), Jurgens2018 (Jurgens et al., 2018) et Su2019 (Su et al., 2019). Ces jeux de données ont été entièrement réannotés par Jiang (2023) selon un schéma fondé sur Teufel (2006), mais adapté et étendu. Les principales différences incluent (i) la fusion des classes "CoCo-" (les résultats de l'article citant sont considérés comme supérieurs aux résultats cités) et "CoCoRes" (comparaison/contraste des résultats), (ii) la fusion des classes "PBas" (les résultats cités sont à la base des travaux de l'article citant) et "PModi" (l'auteur adapte ou modifie un outil ou algorithme, ou des données), et (iii) l'ajout de la classe "Future", reprise de Jurgens (2018), qui caractérise les citations pertinentes pour des travaux futurs.

Afin d'étudier la transférabilité de ces classes à une autre discipline scientifique que la linguistique computationnelle, nous avons constitué un nouveau corpus, PD100cit. Ce corpus est centré sur la maladie du *Pear Decline*, une maladie causée par une bactérie transmise aux poiriers par des insectes et répandue mondialement. Ce cas d'étude permet d'examiner la fiabilité des découvertes citées, issues de sources variées (articles scientifiques, littérature grise), et parfois anciennes, ce qui a permis de s'intéresser à l'évolutivité des connaissances. Dulor (2024) a rassemblé 74 documents pertinents autour du *Pear Decline*, convertis en XML à l'aide de l'outil GROBID (Lopez, 2009). Pour cibler les articles de recherche contenant des découvertes originales, seuls les documents mentionnant explicitement un insecte vecteur ont été conservés. La détection initiale du type de document s'est appuyée sur Zotero, puis a été vérifiée manuellement pour exclure les revues, articles de données, chapitres d'ouvrages, documents de conférence ou autres sources non centrées sur des résultats primaires. Le tri a également conduit à l'exclusion des documents ne comportant pas de citations, ou traitant exclusivement de taxonomie ou de morphologie des insectes, ainsi que des fichiers avec une qualité de structuration XML insuffisante pour les extractions de citations. À l'issue de cette procédure, 24 articles ont été retenus. L'extraction automatique des citations a été réalisée à partir des balises bibliographiques contenues dans les fichiers XML, en utilisant des expressions régulières. Cette étape a permis d'identifier 1 686 citations. Chaque citation a été contextualisée en extrayant la citance (phrase contenant la référence), avec les trois phrases précédentes et les trois phrases suivantes. Parmi ces 1 686 citations, un échantillon de 100 occurrences a été tiré aléatoirement pour une annotation manuelle. Chaque passage de citation a été annoté par le premier auteur selon plusieurs dimensions : la fonction rhétorique ; la fonction biologique ; la polarité ; et le statut attribué au rôle de l'insecte vecteur (confirmé, probable, non confirmé). Par ailleurs, chaque phrase composant le passage de citation a été annotée comme contenant ou non une information pertinente pour la classification rhétorique mentionnée ci-dessus. La procédure d'annotation est détaillée dans notre guide d'annotation.

3.2 Choix des classes

Le corpus PD100cit a été annoté selon la typologie en 11 classes proposée par Jiang (2023), qui catégorise les fonctions rhétoriques des citations dans les textes scientifiques. Cette classification permet notamment d'évaluer la posture argumentative de l'auteur vis-à-vis des sources citées. Par exemple, les classes "support", "weakness" et "contrast/comparison" signalent une opinion explicite (accord, critique, comparaison ou contraste), tandis que "neutral" désigne une mention factuelle

sans prise de position. La table 1 présente des définitions des classes ainsi que des exemples tirés de PD100cit. Ces informations sont également disponibles sur le guide d'annotation.

Classe	Définition	Exemple
Basis	La citation reconnaît les fondements intellectuels du travail actuel.	Based on the information obtained in this and previous studies (14,15 ; A. H. Purcell, unpublished data), pear growers are now aware that pears are the primary reservoir for PYLR in northern California.
CoCoGM	Contraste ou comparaison des objectifs ou méthodes entre l'article citant et l'article cité.	The PCR amplification conditions were the same as proposed by Ghanim et al. [43]
CoCoRes	Comparaison ou contraste des résultats entre l'article citant et l'article cité.	Comparing the detection of grapevine yellows phytoplasma in planthoppers, only 66% of the PCR positives were also positive by enzyme-linked immunosorbent assay (35).
CoCoXY	Comparaison ou contraste entre deux articles cités.	Ullman and Mclean (1986) and Garzo et al. (2012) also observed the same number of teeth on the mandibles of the psyllids C. pyricola and Diaphorina citri respectively, whereas Pollard (1970) found 8 teeth in adults (7 teeth in nymphs) on the mandibles of C. mali.
Future	Mentionne des perspectives ou des travaux futurs.	(Non observée dans PD100cit)
Motivation	Justifie la recherche actuelle par des résultats antérieurs prometteurs ou des enjeux scientifiques.	Regarding the capability of B. nigricornis to transmit CaLsol, previous field work has shown that B. nigricornis can become naturally infected with CaLsol haplotype E (Teresani et al. 2014 ; 2015), so further research to assess the vector efficiency of this psyllid species was needed.
Neutral	Fournit des informations de fond sans positionnement.	The number of protrusions varies among different species, which may be related to the hardness of the leaves of the host plant (Forbes, 1977 ; Rosell et al., 1995 ; Zhao et al. ; Garzo et al., 2012).
Similar	Met en évidence une similarité.	(Non observée dans PD100cit)
Support	Apporte un soutien ou une confirmation.	Sequence similarity values within taxa and divergence between taxa largely confirm the results of previous work (Seemüller et al., 1994 ; Seemüller et al., 1998b).
Usage	Mentionne une méthode, un outil ou des données utilisés dans le travail du citant.	In direct PCR or the first amplification of semi-nested PCR, the universal phytoplasma primers P1/P7 (Deng and Hiruki, 1991 ; Schneider et al., 1995) were used.
Weakness	Critique ou souligne les limites du travail cité.	To compare our results with those papers, we preferred to use the same methodology despite some potential limitations, such as a poor fit.

TABLE 1 – Définitions et exemples des classes rhétoriques

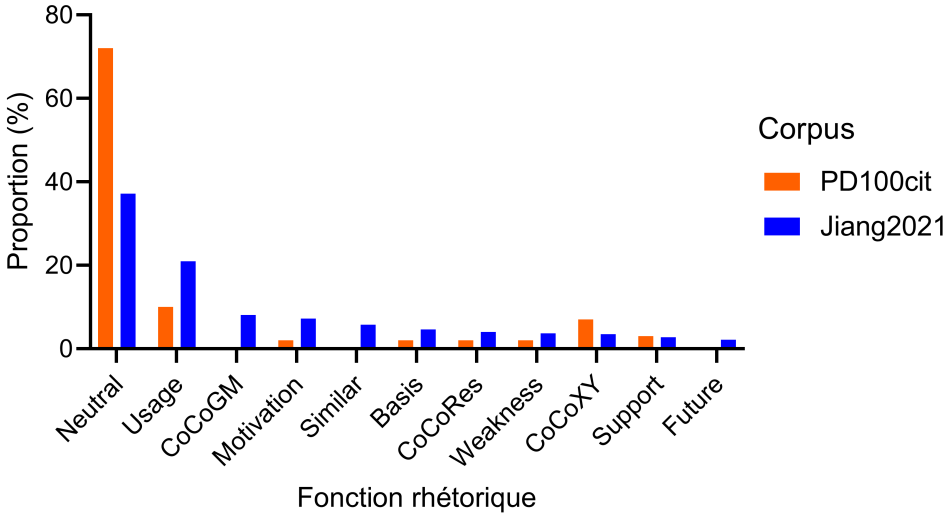

FIGURE 1 – Proportion des fonctions rhétoriques des citations dans les corpus Jiang2021, et 100 citations

Le corpus Jiang2021 contient 118 citations de type "contrast/comparison between 2 cited methods" (CoCoXY), 1 247 "neutral", 123 "weakness", 155 "basis", 703 "usage", 135 "contrast/comparison

in results" (CoCoRes), 243 "motivation", 194 "similar", 272 "contrast/comparison in goals and methods" (CoCoGM), 93 "support", et 73 "future" (Figure 1). Dans le corpus PD100cit, nous trouvons 7 exemples de CoCoXY, 72 "neutral", 2 "weakness", 2 "basis", 10 "usage", 2 CoCoRes, 2 "motivation", 0 "similar", 0 CoCoGM, 1 "support", et 0 "future" (Figure 1). L'absence, ou le petit nombre d'exemples, dans certaines classes résulte de la petite taille du corpus et de spécificités disciplinaires.

3.3 Architecture des modèles

Afin de comparer les performances des modèles entraînés en linguistique informatique sur la classification des fonctions rhétoriques en écologie, nous avons d'abord pris comme référence l'architecture de Jiang et Chen (2023) pour reproduire les résultats de classification sur 11 classes sur le corpus Jiang2021. Leur approche évalue les éléments à représenter (citation, citance, contexte), explore différentes techniques d'agrégation et teste la classification à plusieurs niveaux de complexité (12, 11, 9, 7, et 6 classes). Leur modèle n'étant pas accessible, nous l'avons reconstruit d'après leur article. L'entrée est une séquence alternant la citance et des phrases du contexte droit et gauche, jusqu'à 512 tokens. La citation est masquée par le token CITSEG, dont la représentation est extraite. Une couche de *max-pooling* sur SciBERT produit la représentation du contexte séquentiel. Les représentations sont concaténées, puis envoyées dans perceptron multicouches (MLP) composé de trois couches, dont les couches cachées ont une taille deux fois supérieure. L'attention fixée à 250, mentionnée par Jiang et Chen (2023), reste ambiguë quant à son implémentation exacte. Nous avons interprété cette information, soit comme une réduction directe des dimensions de 768 à 250 avant ou après le *pooling*, soit comme l'ajout d'une couche nn.TransformerEncoderLayer de dimension 250 appliquée à la représentation de CITSEG. Cette architecture n'a pas permis de reproduire les résultats expérimentaux de Jiang et Chen (2023), sans que nous disposions de plus de détails sur l'architecture ou l'accès au code pour reproduire l'expérience.

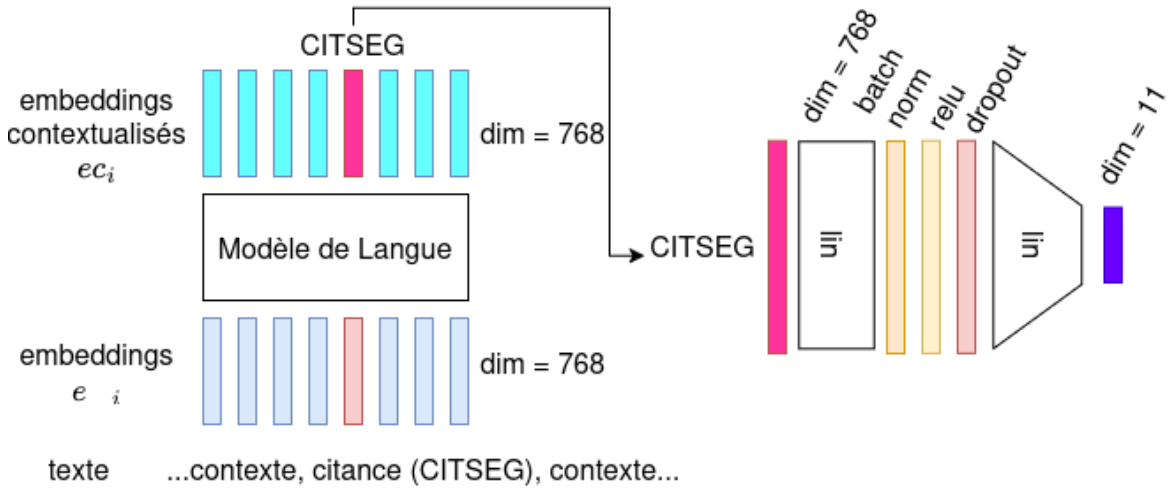

FIGURE 2 – Architecture du modèle de classification rhétorique des citations

Notre propre architecture (Figure 2) repose sur une approche similaire, consistant à extraire la représentation de CITSEG via le modèle de langue, puis à la connecter à un classifieur. Plutôt que de tenter de représenter l'ensemble du contexte, nous supposons que les éléments argumentatifs pertinents

sont condensés dans la représentation de la référence masquée par le token CITSEG, dès lors qu'il est placé dans un contexte structuré (citance et phrases de contextes). Le classifieur est composé d'une couche cachée introduisant de la non-linéarité, de dimensions (768, 768), correspondant à la taille des représentations produites par les modèles de langue. Une normalisation par lot est appliquée pour stabiliser l'apprentissage, suivie d'une activation ReLU. L'enjeu de cette architecture est de spécialiser le modèle de langue afin de projeter les éléments linguistiques discriminants sur la représentation de la citation.

3.4 Paramètres des modèles

Nous avons comparé quatre modèles de langue : SciBERT à des fins de comparaison avec les travaux antérieurs, BioBERT et BioLinkBERT pour leur spécialisation en biologie, et RoBERTa pour sa capacité à modéliser les relations entre tokens. Trois longueurs de fenêtres de contexte ont été explorées (1-1, 2-2, 3-3) afin d'évaluer l'impact de l'information contextuelle. Elles sont notées N-N, où N représente le nombre de phrases avant et après la citance. Les hyperparamètres suivants ont implémentés avec PyTorch 1.10.2+cu102 :

— Reproduction du modèle de Jiang et Chen (2023) : taille de lot = 16, nombre d'époques = 20, graine = 5171, ratio d'échauffement = 0.1, fenêtre de contexte = 2-3, optimiseur = AdamW, coefficient de décroissance des poids = 0.002, répartition = 65% train, 25% validation, 20% test, taux d'apprentissage = 5e-5 (SciBERT) et 5e-4 (MLP).

— Notre modèle : taille de lot = 32, nombre d'époques = 20, graine = 42, 5171, 798, 1965, ratio d'échauffement = 0.2, fenêtre de contexte = 3-3, 2-2, 1-1, optimiseur = AdamW, coefficient de décroissance des poids = 0.002, dropout = 0.5, taux d'apprentissage = 2e-5, répartition = 80% Jiang2021 (train), 20% Jiang2021 (validation), PD100cit (test).

3.5 *Prompt* de GPT

Une instruction, disponible sur github, a été conçu manuellement pour GPT-4 afin de comparer ses performances aux résultats des modèles de langue affiné sur les mêmes longueurs de fenêtres (3-3, 2-2 et 1-1). Le modèle GPT-4 a été choisi pour sa capacité à généraliser sur des tâches complexes. Le *prompt* comprend une formulation de la tâche, les définitions des classes extraites du guide d'annotation, et un exemple par classe. Les exemples ont été choisis de sorte à éviter les recouvrements entre classes afin que le modèle puisse distinguer celles-ci. La tâche a été formulée de la façon suivante : "you are a classifier that assigns the references in a scientific article passage to a rhetorical class. The rhetorical class represents the argumentation role of the reference. The reference will be denoted by (CITSEG)". Ainsi, pour chaque passage contenant une citation annotée (CITSEG), GPT-4 doit générer une classe parmi celles proposées dans le *prompt*.

4 Résultats

4.1 Fine-tuning des modèles SciBERT, RoBERTa, BioBERT et BioLinkBERT

Malgré nos tentatives de reproduction de l'architecture de Jiang et Chen (2023), les résultats expérimentaux restent très inférieurs à ceux rapportés dans leur article. Nos meilleurs résultats atteignent 10.13% de F1 macro et 25.95% de F1 pondérée, contre un F1 maximal de 66.16% sur onze classes selon Jiang et Chen (2023). Nos scores ont été obtenus en ajoutant une couche nn.TransformerEncoderLayer sur la représentation de CITSEG. Nous avons également testé des réductions de dimension avant ou après le *pooling*, sans observer de différence notable. Ces modèles présentent un surapprentissage. Ces écarts de performance peuvent s'expliquer par les ambiguïtés dans la description de l'architecture, en particulier concernant la manière dont la dimension d'attention fixée à 250 est implémentée. Un accès au code serait nécessaire pour permettre une reproduction fidèle.

Avec notre architecture appliquée au corpus Jiang2021, SciBERT présente la meilleure F1 pondérée moyenne (67.18%) avec un écart-type de 2.45 (Table 1), tandis que BioLinkBERT atteint la meilleure F1 macro moyenne (52.34%) avec un écart-type de 2.38, toutes deux en longueur de fenêtre 3-3 (Table 1). Le meilleur score de validation est toutefois atteint par BioBERT sur la longueur de fenêtre 2-2 avec une F1 macro de 58.29% et une valeur F1 pondérée de 70.46% (Table 2). Les écarts moyens entre modèles sont serrés, entre 0.66 et 1.33 points pour la valeur F1 pondérée et 0.35 et 1.19 points pour la F1 macro sur la fenêtre 3-3.

	Contexte 1-1		Contexte 2-2		Contexte 3-3	
Modele	F1_macro	F1_pondérée	F1_macro	F1_pondérée	F1_macro	F1_pondérée
BioBERT	50.69 (3.55)	65.31 (2.34)	51.92 (3.91)	65.87 (2.74)	51.64 (3.21)	66.43 (2.00)
BioLinkBERT	51.57 (3.11)	66.17 (2.13)	51.66 (2.39)	66.10 (2.17)	**52.34** (2.38)	66.52 (2.25)
RoBERTa	50.45 (3.85)	65.21 (2.66)	51.12 (1.82)	65.41 (2.31)	51.15 (3.06)	65.85 (2.09)
SciBERT	49.89 (2.74)	65.38 (1.99)	52.17 (2.98)	66.70 (2.22)	51.99 (3.48)	**67.18** (2.45)

TABLE 2 – Valeurs F1 macro et F1 pondérée moyennes sur les quatre graines et écart-type entre parenthèse par fenêtre de contexte et par modèle sur l'ensemble de validation corpus Jiang2021

Pour notre analyse de la qualité du modèle transféré sur le corpus PD100cit, nous considérons la métrique F1 pondérée, car elle ajuste les scores en fonction de la distribution des classes, contrairement à la F1 macro qui accorde trop de poids aux classes rares. Sur PD100cit, la meilleure F1 pondérée moyenne est obtenue avec le modèle BioBERT sur la longueur de fenêtre 3-3 avec une F1 pondérée de 78.08% significativement supérieure aux résultats obtenus sur le corpus Jiang2021 (écart-type de 4.91) (Figure 3).

Les performances varient selon les graines par longueur de fenêtre. Bien que la longueur de fenêtre 3-3 donne les meilleurs scores, la longueur de fenêtre 2-2 est la plus stable, avec une valeur F1 pondérée entre 68.46% et 78.86% sur les quatre modèles. La variabilité des scores pour la longueur de fenêtre 3-3, (58.46% à 82.91%) sur l'ensemble des modèles pourrait résulter de l'ajout excessif d'informations peu pertinentes. Les résultats avec la fenêtre 1-1 sont moins stables, en particulier pour RoBERTa, avec un écart de 20.12 points en valeur de F1 pondérée. BioBERT étant le modèle le plus constant sur les différentes graines sur PD100cit, nous approfondissons dans la suite de l'article

Modèle	Graine	Contexte 1-1		Contexte 2-2		Contexte 3-3	
		Best_F1m	Best_F1w	Best_F1m	Best_F1w	Best_F1m	Best_F1w
BioBERT	1965	47.66	63.59	47.93	64.99	46.63	64.81
	42	46.84	63.99	49.77	64.83	51.65	65.11
	5171	55.31	69.33	**58.29**	**70.46**	55.50	69.82
	798	52.97	64.32	51.69	63.20	52.78	65.98
BioLinkBERT	1965	49.68	65.15	50.48	65.56	50.85	65.24
	42	48.04	64.00	48.57	63.03	51.80	65.12
	5171	56.29	69.68	54.98	69.03	**56.36**	**70.41**
	798	52.27	65.87	52.60	66.78	50.36	65.30
RoBERTa	1965	47.11	62.21	49.83	64.57	50.30	65.18
	42	46.16	63.56	50.60	64.26	48.19	64.29
	5171	54.78	69.23	54.22	69.35	**56.28**	**69.43**
	798	53.77	65.86	49.82	63.47	49.84	64.49
SciBERT	1965	48.11	64.24	49.67	65.54	48.10	65.07
	42	46.43	63.16	49.40	65.25	49.52	65.56
	5171	53.24	68.46	56.78	70.55	**57.03**	**71.29**
	798	51.78	65.65	52.84	65.47	53.30	66.82

TABLE 3 – Meilleurs scores F1 macro (Best_F1m) et F1 pondérée (Best_F1w) par modèle, graine et longueur de fenêtre sur l'ensemble de validation corpus Jiang2021

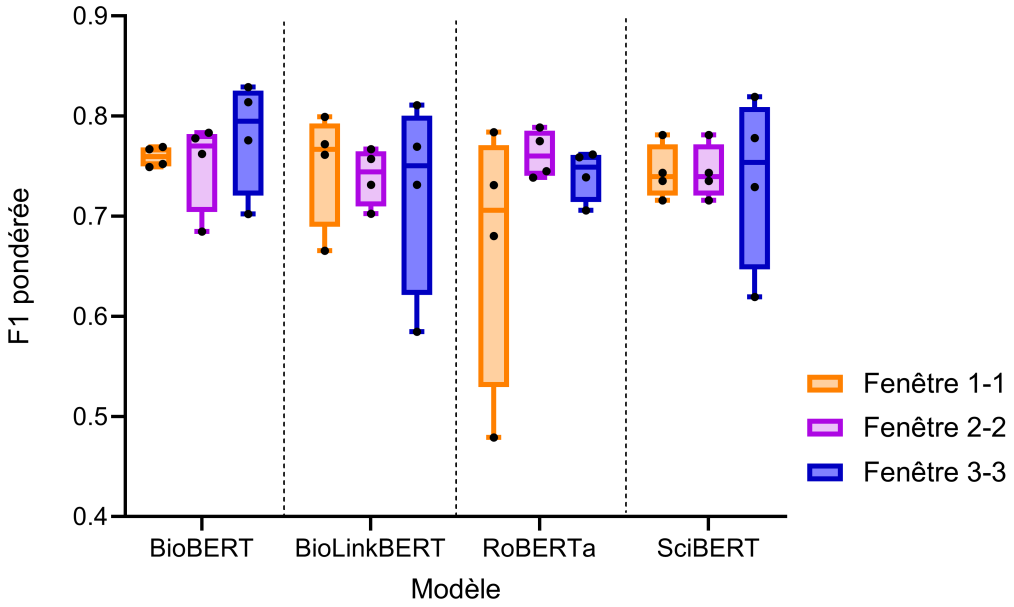

FIGURE 3 – Valeurs de F1 pondérée sur PD100cit par longueur de fenêtre et par modèle sur les quatre graines

les analyses sur ce modèle pour la longueur de fenêtre 3-3 pour laquelle les meilleures performances ont été obtenues.

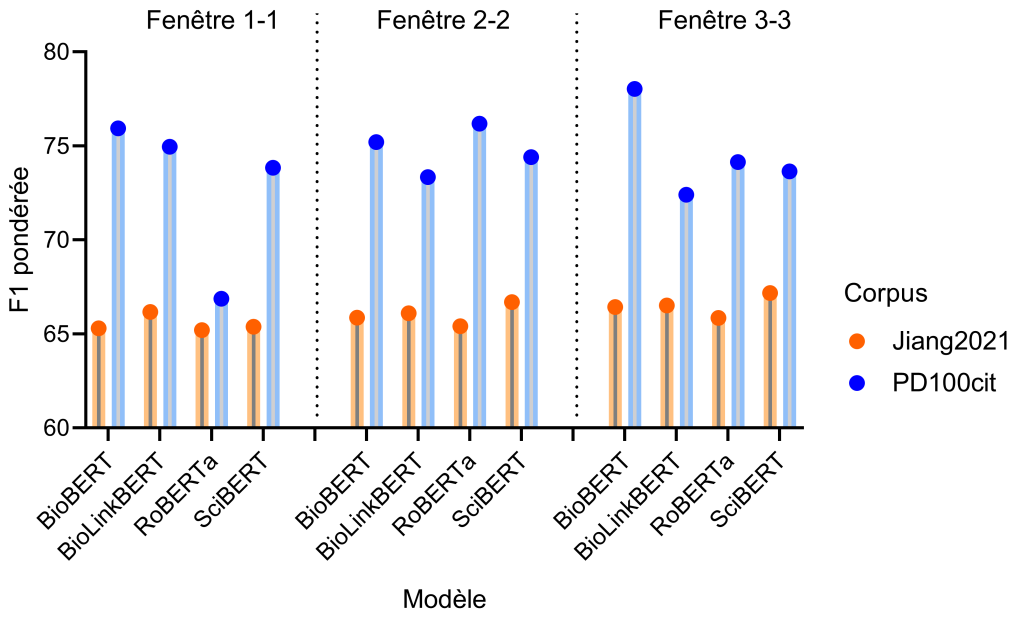

FIGURE 4 – Valeurs de F1 pondérée moyenne par modèle et par fenêtre sur PD100cit

Les valeurs de F1 pondérée sont plus élevées sur PD100cit que sur Jiang2021 (Figure 4), ce que nous expliquons par la répartition entre classes, 72% des citations de PD100cit étant dans la classe "neutral" contre 37.16% pour Jiang2021. Les meilleures performances moyennes sur PD100cit sont obtenues sur la classe “neutral” avec une précision moyenne de 88.60% et un rappel moyen de 83.68% (Table 3), tandis que sur Jiang2021, la classe “neutral” obtient une précision moyenne de 74.20% et un rappel moyen de 68.62% (Table 4). Si le modèle prédisait systématiquement la classe “neutral” sur PD100cit, la précision serait de 72%, le rappel de 100% et la F1 83.72%, inférieures aux scores obtenus. Ainsi, les performances obtenues pour la classe "neutral" ne dépendent pas seulement de sa forte proportion, mais aussi d’autres facteurs comme des éléments linguistiques spécifiques aux articles de biologie de PD100cit. Les spécificités du domaine pourraient également expliquer la bonne classification de la classe “usage”, qui obtient une F1 de 74.09%. Cette classe se caractérise par un vocabulaire spécifique, notamment des mentions de “primers” pour les techniques d’identification, comme illustré par la citation : “... two different pairs of ribosomal primers used were : fPD/r0 1 (CITSEG) and AP3/ AP5 (Firrao et al., 1994) respectively.”. De même, toutes les citations de la classe “support” ont été prédites, probablement en raison d’expressions clés comme “confirm the results of”. La classe CoCoRes est également bien prédite par le modèle malgré le petit nombre d’instances. Toutefois, les citations CoCoXY sont confondues avec celles de la classe majoritaire “neutral” comme le montre la matrice de confusion (Figure 5). Les classes “weakness” et “neutral” peuvent également être difficiles à distinguer, comme le montre cet exemple : “More than 33 psyllid species infest cultivated pear trees around the world with the taxonomic status of most of them still not clear ((CITSEG) ; Luo et al. 2012)”. Cette citation, que nous avons classée en "weakness", peut être interprétée de deux manières : soit pour souligner l’incertitude, soit comme une reprise des dires des auteurs, ce qui la classerait en “neutral”.

Pour analyser les confusions entre classes, nous examinons les paires d’occurrences de *true label* (TL),

Classe	Nb citations	P	R	F1
Neutral	72	88.60	83.68	85.59
Usage	10	82.64	67.5	74.09
CoCores	2	75	62.5	66.67
Motivation	2	39.58	37.5	35.0
Basis	2	0.00	0.00	0.00
Weakness	2	83.33	62.5	68.34
Support	3	64.58	100.00	75.59
CoCoXY	7	43.61	60.71	47.52

TABLE 4 – Précision, Rappel, F-mesure et nombre de citations moyens (en %) par classe sur le corpus PD100cit en test avec BioBERT affiné en fenêtre 3-3

Classe	P	R	F1
Neutral	74.20	68.62	70.75
Usage	73.52	79.76	76.32
CoCores	65.08	61.87	63.10
Motivation	48.86	58.47	51.96
Basis	55.40	53.99	54.12
Weakness	47.48	46.48	45.54
Support	39.49	27.70	35.46
CoCoXY	27.09	18.97	21.29
CoCoGM	60.66	66.36	63.17
Similar	61.05	62.76	61.77
Future	68.83	60.78	64.41

TABLE 5 – Précision, Rappel et F-mesure moyens (en %) par classe sur l'ensemble de validation corpus Jiang2021 avec BioBERT affiné en fenêtre 3-3

Top1 (T1), Top2 (T2) et Top3 (T3), en excluant la classe “neutral” (Figure 5). La figure 5 présente les résultats obtenus avec BioBERT sur la graine 42, configuration pour laquelle nous avons obtenu la meilleure F1 pondérée (82,91%) sur PD100cit. L’objectif est d’identifier les classes fréquemment associées, où le modèle a montré des hésitations. Les paires les plus courantes sont : CoCoXY et “support”, "usage" et CoCoXY, “basis” et CoCoXY, CoCoRes et CoCoXY. Ces confusions peuvent résulter de similarités linguistiques, d’ambiguïtés dans l’objet de comparaison ou d’un recouvrement sémantique des classes.

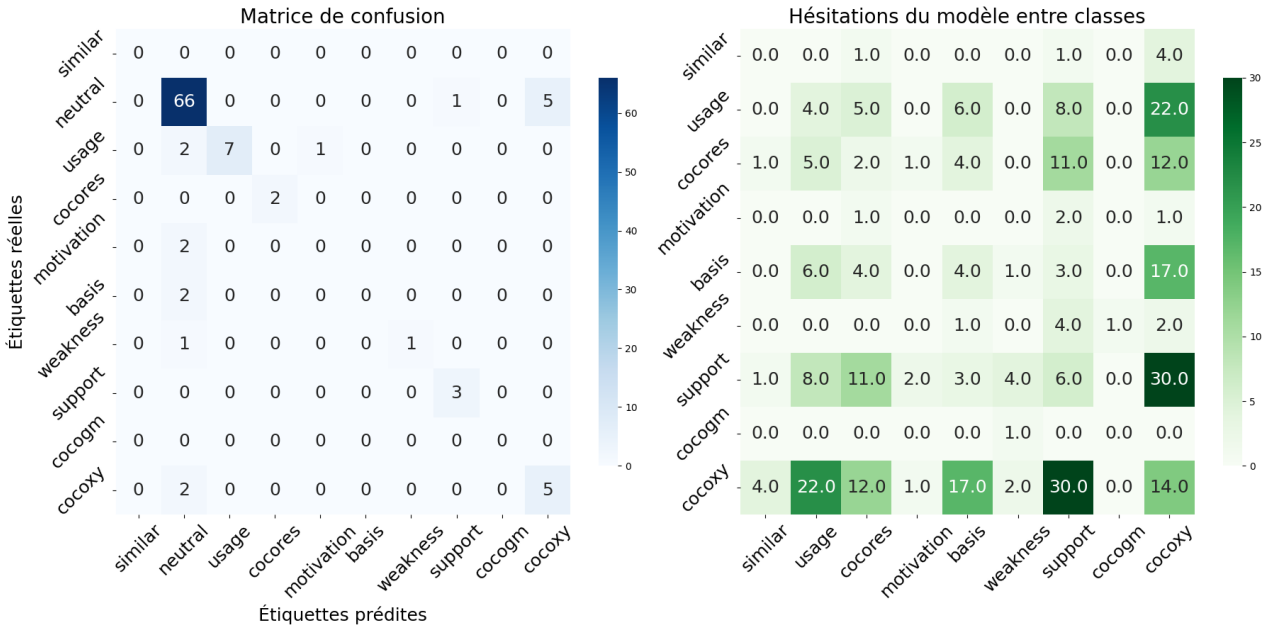

FIGURE 5 – Matrice de confusion et hésitations du modèle entre classes (hors "neutral") avec BioBERT affiné en fênetre 3-3 sur la graine 42

4.2 Résultats du *prompt* de GPT

Les performances de GPT-4 pour la classification des citations de PD100cit en fonction des longueurs de fenêtres 3-3, 2-2 et 1-1 sont respectivement 37.38%, 35.09% et 31.98% de valeur F1 pondérée et 18.18%, 19.07% et 23.14% en F1 macro. En réduisant la longueur de la fenêtre, la valeur F1 pondérée diminue tandis que la F1 macro augmente. Une petite fenêtre semble favoriser certaines petites classes comme “support”, “motivation” et CoCoRes aux dépens des classes plus grandes comme “neutral” et “usage” (Table 7).
Comparé aux meilleurs scores de BioBERT avec une longueur de fenêtre 3-3, GPT-4 affiche une baisse de 45.53 points de valeur F1 pondérée. Cette baisse est marquée par une mauvaise prédiction de la classe fréquente “neutral” (F1=41.76%). La classe la plus prédite est “support”, avec un rappel est presque dix fois supérieur à la précision en fenêtre 3-3. La classe "weakness" est toujours retrouvée. La deuxième classe la mieux prédite est "usage", avec une mesure de F1 de 45.71% en longueur de fenêtre 3-3, ce qui reste très inférieur aux performances de BioBERT affiné.

Des expériences complémentaires de GPT-4 avec affinage et d’autres modèles génératifs seront utiles pour confirmer les moindre performances en classification rhétorique par rapport aux modèles encodeurs bidirectionnels de type BERT.

Classe	Nombre de citations	Contexte 3-3			Contexte 2-2			Contexte 1-1		
		P	R	F1	P	R	F1	P	R	F1
Neutral	72	100.00	26.39	41.76	94.44	23.61	37.82	100.00	19.44	32.56
Usage	10	32.00	80.00	45.71	29.63	80.00	43.24	28.12	90.00	42.86
CoCores	2	11.11	50.00	18.18	12.50	50.00	20.00	28.57	100.00	44.44
Motivation	2	0.00	0.00	0.00	50.00	100.00	66.67	66.67	100.00	80.00
Basis	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Weakness	2	100.00	100.00	100.00	40.00	100.00	57.14	40.00	100.00	57.14
Support	3	6.90	66.67	12.50	13.04	100.00	23.08	11.54	100.00	20.69
CoCoXY	7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE 6 – Précision, Rappel et F-mesure (en %) par classe et par longueur de fenêtre par *prompt* de GPT-4

5 Conclusion

Dans cet article, nous avons présenté PD100cit, un corpus annoté pour la classification rhétorique des citations en écologie, disponible avec son guide d'annotation et les codes de classification. Cette étude confirme les capacités de transfert des modèles de langue affinés en linguistique computationnelle vers l'analyse d'interactions complexes au sein de systèmes biologiques pour la classification des fonctions rhétorique des citations. Nos résultats montrent que les modèles de langue affinés surpassent le *prompt* de GPT-4 sans affinage, soulignant les avantages de l'apprentissage sur des corpus de linguistique computationnelle disponibles vers des domaines moins dotés. Toutefois, davantage de *prompts* devraient être expérimentés afin de mesurer l'effet de la variabilité des *prompts* sur les résultats. Nous avons notamment exploré la sensibilité des modèles à la longueur des contextes. L'usage de la typologie des fonctions rhétoriques des citations utilisée en linguistique computationnelle a pu être étendu aux interactions biologiques, bien que des ajustements resteront nécessaires pour évaluer la véracité des découvertes citées. Les résultats du transfert des modèles de langues affinés pour la classification des citations sur la maladie *Pear Decline* encouragent l'étude de la fiabilité des articles scientifiques dans le contexte plus général des maladies émergentes à forte incidence en santé des plantes, en santé animale et humaine.

Ces analyses devront être confirmées à un plus large corpus couvrant d'autres cas d'études biologiques. La typologie pourrait nuancer les citations négatives et fusionner les classes moins pertinentes pour estimer la fiabilité. Les modèles pourront inclure des métadonnées comme les sections et titres des articles (Zhang et al., 2022). D'autres approches génératives pourront être explorées telles que DeepSeek (Bi et al., 2024) en étendant les instructions par l'ajout de contraintes négatives.

Remerciements

Nous remercions la plateforme de bioinformatique MIGALE de l'INRAE (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi : 10.15454/1.5572390655343293E12) pour son aide et/ou la mise à disposition de ressources de calcul et/ou de stockage. Ces travaux ont été initiés grâce à des collaborations au sein du projet BEYOND (contrat ANR n° 20-PCPA-0002). Anne-Sophie Foussat est actuellement en thèse et bénéficie d'un financement de la Direction Générale INRAE.

Références

BELTAGY I., LO K. & COHAN A. (2019). SciBERT : A Pretrained Language Model for Scientific Text. 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)*, p. 3613–3618 : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1371.

BI X., CHEN D., CHEN G., CHEN S., DAI D., DENG C., DING H., DONG K., DU Q., FU *et al.* (2024). DeepSeek LLM : Scaling Open-Source Language Models with Longtermism. *arXiv*, **2401.02954**. DOI : 10.48550/arXiv.2401.02954.

BORNMANN L. (2019). Citation Concept Analysis (CCA) – A New Form of Citation Analysis Revealing the Usefulness of Concepts for Other Researchers Illustrated by Two Exemplary Case Studies Including Classic Books by Thomas S. Kuhn and Karl R. Popper. *Scientometrics*, **122**(2), 1051–1074. DOI : 10.1007/s11192-019-03326-2.

COHAN A., AMMAR W., VAN ZUYLEN M. & CADY F. (2019). Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 3586–3596 : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1361.

DULOR M. (2024). Qualification de la vraisemblance des informations publiées dans la littérature scientifique - Étude d'un cas concret : 'Candidatus Phytoplasma pyri' et Pear decline. Mémoire de master, Université Paris 10 - Nanterre & INRAE. https://asnr.hal.science/hal-04662925.

GARZONE M. & MERCER R. E. (2000). Towards an Automated Citation Classifier. In *Advances in Artificial Intelligence*, volume 1822, p. 337–346. Springer Berlin Heidelberg. DOI : 10.1007/3-540-45486-1_28.

HARWOOD N. (2009). An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, **41**(3), 497–518. DOI : 10.1016/j.pragma.2008.06.001.

HASSAN S.-U., IQBAL S., IMRAN M., ALJOHANI N. R. & NAWAZ R. (2018). Mining the Context of Citations in Scientific Publications. In *Maturity and Innovation in Digital Libraries*, volume 11279, p. 316–322. Springer International Publishing. DOI : 10.1007/978-3-030-04257-8_32.

HU G. & WANG G. (2014). Disciplinary and ethnolinguistic influences on citation in research articles. *Journal of English for Academic Purposes*, **14**, 14–28. DOI : 10.1016/j.jeap.2013.11.001.

HYLAND K. (1999). Academic attribution : citation and the construction of disciplinary knowledge. *Applied Linguistics*, **20**(3), 341–367. DOI : 10.1093/applin/20.3.341.

JARAUSCH W., JARAUSCH B., FRITZ M., RUNNE M., ETROPOLSKA A. & PFEILSTETTER E. (2019). Epidemiology of European stone fruit yellows in Germany : the role of wild *Prunus spinosa*. *European Journal of Plant Pathology*, **154**(2), 463–476. DOI : 10.1007/s10658-019-01669-3.

JIANG X. & CHEN J. (2023). Contextualised segment-wise citation function classification. *Scientometrics*. DOI : 10.1007/s11192-023-04778-3.

JURGENS D., KUMAR S., HOOVER R., MCFARLAND D. & JURAFSKY D. (2018). Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics*, **6**, 391–406.

KOLOVEAS P., CHATZOPOULOS S., VERGOULIS T. & TRYFONOPOULOS C. (2025). Can LLMs Predict Citation Intent ? An Experimental Analysis of In-context Learning and Fine-tuning on Open LLMs. DOI : 10.48550/arXiv.2502.14561.

KUNNATH S. N., PRIDE D. & KNOTH P. (2023). Prompting Strategies for Citation Classification. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, p. 1127–1137 : ACM. DOI : 10.1145/3583780.3615018.

KUNNATH S. N., STAUBER V., WU R., PRIDE D., BOTEV V. & KNOTH P. (2022). ACT2 : A multi-disciplinary semi-structured dataset for importance and purpose classification of citations. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 3398–3406.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240. DOI : 10.1093/bioinformatics/btz682.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach. DOI : 10.48550/arXiv.1907.11692.

LOPEZ P. (2009). GROBID : Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL)* : Springer.

MORAVCSIK M. J. & MURUGESAN P. (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science*, **5**(1), 86–92. DOI : 10.1177/030631277500500106.

SWALES J. (1986). Citation Analysis and Discourse Analysis. *Applied Linguistics*, **7**(1), 39–56. DOI : 10.1093/applin/7.1.39.

TEUFEL S., SIDDHARTHAN A. & TIDHAR D. (2006). Automatic Classification of Citation Function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*, p. 103–110 : Association for Computational Linguistics. DOI : 10.3115/1610075.1610091.

TEUFEL S., SIDDHARTHAN A. & TIDHAR D. (2009). An Annotation Scheme for Citation Function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, p. 80–87 : Association for Computational Linguistics. DOI : 10.3115/1654595.1654612.

WAN X. & LIU F. (2014). Are all literature citations equally important ? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology*, **65**(9), 1929–1938. DOI : 10.1002/asi.23083.

YASUNAGA M., LESKOVEC J. & LIANG P. (2022). LinkBERT : Pretraining Language Models with Document Links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8003–8016 : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.551.

ZHANG Y., ZHAO R., WANG Y., CHEN H., MAHMOOD A., ZAIB M., ZHANG W. E. & SHENG Q. Z. (2022). Towards Employing Native Information in Citation Function Classification. *Scientometrics*, **127**(11), 6557–6577. DOI : 10.1007/s11192-021-04242-0.

ZHENG Q. & LI B. (2022). Disciplinary and Generic Variation of Citation Use in Research Articles. *International Journal of English for Academic Purposes : Research and Practice*, (Spring), 41–59. DOI : 10.3828/ijeap.2022.4.