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

Vers des interfaces favorisant l'engagement critique des utilisateurs : un prototype utilisant RAG

Petra Dadić[1] Liana Ermakova[1]
(1) HCTI, 20 rue Duquesne, 29200 Brest, France
dadic@univ-brest.fr,liana.ermakova@univ-brest.fr

RÉSUMÉ

Les chatbots à LLM sont devenus omniprésents, mais produisent parfois des informations trompeuses (« hallucinations »), souvent difficiles à repérer pour les utilisateurs. Cet article de démonstration présente une interface prototype conçue pour aider les utilisateurs à identifier et vérifier les informations cruciales dans le contenu généré par l'IA. Dans le cadre de la génération augmentée de recherche (RAG), l'interface met en évidence les informations clés et fournit un accès en temps réel aux sources de soutien ou contradictoires. Nous avons mené une étude avec 80 participants pour recueillir des retours et affiner la conception, en mettant l'accent sur l'amélioration de la source d'information et de la confiance des utilisateurs. Cet article montre comment l'interface peut aider à repérer la désinformation et à améliorer l'usage des LLM pour la recherche d'informations scientifiques.

ABSTRACT

Interfaces for Supporting Critical User Engagement : A Prototype Using RAG

LLM chatbots are now part of daily life, but can still generate hard-to-spot misleading information ("hallucinations"). This demo paper presents a prototype interface designed to help users identify and verify critical information in AI-generated content. Within the Retrieval-Augmented Generation (RAG) setting, the interface highlights key information and provides real-time access to supporting or contradictory sources. We conducted a study with 80 participants to gather feedback and refine the design, focusing on improving information sourcing and user trust. This paper shows how interface design can help users detect misinformation and improve LLMs' usefulness in academic search.

MOTS-CLÉS : Interface graphique, RAG, esprit critique.

KEYWORDS: Search interfaces, RAG, critical thinking.

ARTICLE : **Soumis à** CORIA 2025.

1 Introduction

Ces dernières années, l'essor spectaculaire de l'IA a démocratisé les chatbots, devenus des outils quotidiens, mais les grands modèles de langage (LLM) produisent parfois des informations fausses mais crédibles (« hallucinations ») (Ateia & Kruschwitz, 2023; Ermakova *et al.*, 2023; Ji *et al.*, 2023). Les LLM génèrent aujourd'hui des textes très proches de l'écriture humaine, rendant la détection d'erreurs (Pride *et al.*, 2023) et l'évaluation de la fiabilité des contenus, voire de leur existence, de plus en plus difficiles. Ce phénomène, répandu (Press *et al.*, 2024; Ghanem *et al.*, 2024), affecte

même les modèles les plus performants. Par exemple, ChatGPT ne fournit des références exactes ou partiellement exactes que dans 50% des cas, mais seuls 10% des documents cités existent réellement (Zuccon *et al.*, 2023). Une étude menée dans le cadre de la tâche 3 de CLEF 2024 SimpleText (Ermakova *et al.*, 2024b,a) , dédiée à la simplification automatique de textes scientifiques, a permis de mesurer ce contenu erroné. Elle s'est appuyée sur l'identification de phrases générées ne partageant aucun mot avec le texte source. Ces contenus parasites étaient fréquents : 47% des soumissions en comportaient dans au moins 10% des cas, 39% dans 20% des cas, et 19% dans 50% des cas.

Ces problématiques montrent la nécessité d'éviter que des informations incorrectes ne soient présentées à l'utilisateur. Pour cela, plusieurs méthodes ont été développées (Gao *et al.*, 2022; Thoppilan *et al.*, 2022; Chen *et al.*, 2024; Yan *et al.*, 2024), notamment des techniques de vérification des faits fondées sur la génération augmentée de récupération (Retrieval Augmented Generation, ou RAG) (Lewis *et al.*, 2021). Celles-ci utilisent des sources externes pour enrichir le contexte des modèles de langage, améliorant ainsi la qualité des réponses. Cette approche est particulièrement précieuse en recherche académique (An *et al.*, 2024; Ali *et al.*, 2024; Yin *et al.*, 2025; Wang *et al.*, 2024a), où transparence et citation des sources sont essentielles. Certains outils, comme Scopus AI [1], combinent recherche d'information et synthèse de contenu, mais leur accès est souvent restreint à des bases de données propriétaires, et ils ne sont généralement pas libres d'accès. Les travaux existants sur les systèmes RAG dotés d'interfaces se concentrent le plus souvent sur la compréhension technique de ces systèmes (Wang *et al.*, 2024b), plutôt que sur l'accompagnement des utilisateurs dans leurs recherches ou la vérification rapide d'informations.

Pour sensibiliser les utilisateurs aux risques de désinformation, les recherches sur l'influence de la détection d'erreurs dans les contenus générés sur la confiance en IA suivent généralement deux grandes approches. La première vise à intégrer des indices d'incertitude dans les réponses peu fiables pour aider les utilisateurs à repérer les erreurs potentielles (Belosevic & Buschmeier, 2024; Zhou *et al.*, 2024). La seconde explore l'effet de différents modes de présentation, comme la mise en valeur visuelle ou l'affichage des sources, sur le niveau de confiance accordé aux contenus générés (Papenmeier *et al.*, 2022; Leiser *et al.*, 2024). Des expérimentations ont ainsi testé les préférences des utilisateurs en matière d'explications à travers des captures d'écran (Papenmeier *et al.*, 2022) ou des enregistrements d'interactions avec des interfaces de LLM (Leiser *et al.*, 2024). Toutefois, ces études reposaient sur des déclarations auto-rapportées et non anonymes, limitant leur fiabilité. Elles fournissaient peu d'informations précises sur les sources citées, privilégiant les moyens de mettre en valeur certaines parties du texte généré. L'une d'elles impliquait un nombre réduit de participants et a pu négliger certains comportements ou usage.

Cette étude vise à combler des lacunes des approches actuelles en proposant une interface simple et intuitive, aidant à identifier et vérifier rapidement les informations clés d'un texte généré grâce aux sources associées. Plus précisément, nous cherchons à répondre à la question suivante : « Dans le contexte du RAG, comment aider les utilisateurs à comprendre d'où provient l'information et à repérer d'éventuelles erreurs ou "hallucinations" dans le contenu généré par l'IA ? »

Face aux limites structurelles des modèles actuels, nous privilégions une approche socio-technique en développant des interfaces et outils d'accompagnement adaptés. Nous ciblons l'environnement utilisateur des LLMs, notamment les modèles ouverts comme LLaMA-2 (Touvron *et al.*, 2023). Le système développé vise à renforcer la confiance des utilisateurs envers les contenus générés par l'IA grâce à une interface transparente et intuitive, permettant d'explorer les sources, d'évaluer la fiabilité des affirmations et d'interagir avec l'IA de manière plus éclairée.

1. https://elsevier.libguides.com/Scopus/ScopusAI

Dans le cadre de cette démonstration, nous avons intégré l'API d'Elsevier API[2] afin de concevoir un prototype dédié à la recherche académique assistée par IA, dans un environnement fondé sur le principe du RAG. Le système a été pensé pour renforcer la transparence et la confiance, en affichant des sources académiques, en surlignant les passages pertinents et en proposant des réponses qui tiennent compte du contexte. Le moteur de recherche d'Elsevier est utilisé pour extraire des articles scientifiques, des publications académiques, et des revues à comité de lecture. Ce moteur est utilisé pour chercher des sources qui répondent directement aux requêtes des utilisateurs, telles que « Quelles sont les dernières avancées de l'IA dans le domaine de la santé ? » ou « Quel est l'impact de certains algorithmes sur l'analyse des données ? »

Pour valider les besoins des utilisateurs, et afin d'évaluer les premiers éléments de l'interface dans un environnement pseudo-RAG, nous avons mené une étude contrôlée auprès de 80 participants, représentant des utilisateurs types de systèmes d'information scientifique — des personnes ayant un niveau d'éducation et un parcours professionnel avancés (ano, 2025). L'analyse des résultats a permis de mieux comprendre les attentes des utilisateurs, et de dégager de nouvelles exigences fonctionnelles. Le prototype a ainsi été conçu en tenant compte de cette étude, avec un accent particulier sur l'expérience utilisateur, les préférences exprimées et les retours reçus :

— Nous avons conçu une interface facilitant la compréhension des thématiques abordées grâce à l'intégration d'éléments d'aide visuelle adaptés.
— Les composants de l'interface ont été ajustés pour mieux correspondre aux préférences des utilisateurs, avec la possibilité de soumettre librement des suggestions d'amélioration.
— Nous avons renforcé la vérifiabilité des affirmations générées en permettant un aperçu immédiat du document source ayant contribué à leur élaboration.
— Pour garantir des sources fiables, nous avons choisi de nous concentrer sur des publications scientifiques, en écartant les contenus pouvant prêter à confusion.
— Le système a été intégré dans un cadre RAG réel, en connectant l'interface à l'API d'Elsevier, remplaçant ainsi les données générées artificiellement dans l'étude précédente (ano, 2025).
— L'évaluation repose sur l'analyse des desiderata et des exigences identifiées.

L'article s'organise ainsi : Section 2 présente le test d'utilisabilité (méthodologie, participants, critères d'évaluation) ; Section 3 décrit la conception du prototype selon les retours utilisateurs, l'aperçu du système et ses aspects techniques ; Section 4 illustre ses fonctionnalités via des scénarios ; Section 5 en évalue l'ergonomie et la performance ; Section 6 conclut sur les résultats, limites et perspectives.

2 Test d'utilisabilité

Afin de mieux cerner les besoins des utilisateurs et de valider les exigences fonctionnelles liées au développement de notre système, nous avons conçu une interface web permettant aux participants de tester le modèle proposé (ano, 2025). L'anonymat des participants a été respecté, favorisant ainsi des retours sincères. L'objectif principal était de permettre aux utilisateurs d'interagir avec un chatbot dans un environnement pseudo-RAG contrôlé, afin d'analyser leurs comportements et impressions (voir Fig. 1). L'étude visait notamment à observer comment différents éléments d'interface, comme la mise en forme et l'affichage des sources, influencent la confiance accordée aux informations fournies, tout en recueillant les perceptions générales des utilisateurs vis-à-vis de ce type d'interface conversationnelle. À l'issue de cette analyse, nous avons conçu un prototype fonctionnel,

2. https://dev.elsevier.com/

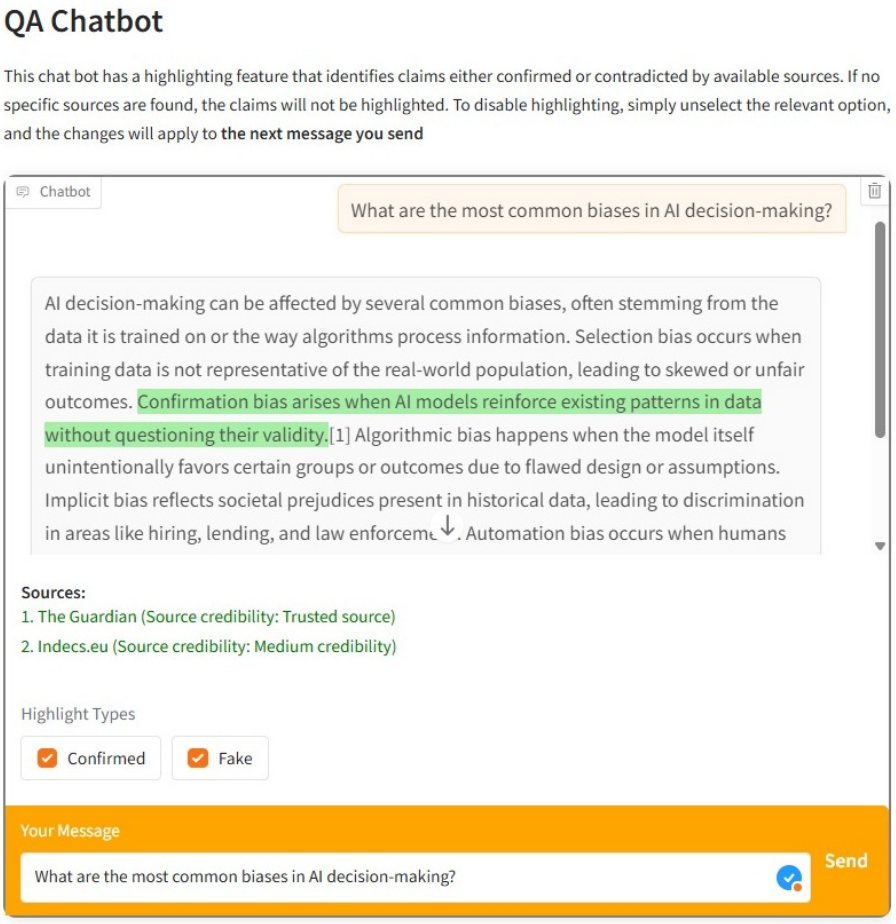

FIGURE 1 – Interface initiale : référencement des sources via RAG avec mise en évidence des sources

intégrant les retours recueillis au cours de l'étude. L'expérimentation s'est déroulée en trois temps : un questionnaire préliminaire, une phase d'interaction avec le chatbot, puis un questionnaire final. L'enquête initiale visait à recueillir des données sociodémographiques ainsi qu'à évaluer l'expérience préalable des participants avec les outils d'IA. Durant la phase d'interaction, pour reproduire certaines limites observées dans les systèmes actuels, les participants ont échangés avec un chatbot incluant volontairement des réponses exactes et d'autres erronées. Enfin, en utilisant une échelle de Likert (Likert, 1932) et des questions ouvertes, le questionnaire post-expérience a permit de recueillir les impressions des utilisateurs sur les éléments d'interface et sur leur expérience globale.

80 participants, majoritairement de France et de Croatie, ont été recrutés via des listes de diffusion académiques et des réseaux sociaux. Représentant des usagers typiques des systèmes d'information scientifique, ils présentaient une répartition hommes-femmes équilibrée. 35 avaient entre 25 et 34 ans. La plupart détenaient un diplôme de niveau master et avaient déjà utilisé des chatbots basés sur l'IA, avec, dans l'ensemble, des expériences jugées positives.

La majorité des participants (71%) ont trouvé l'interface facile à utiliser, tandis que 8% ont rencontré des difficultés. Les avis étaient plus partagés sur les scores de confiance, l'affichage des sources et le surlignage, avec des perceptions divisées sur la fiabilité. Malgré tout, l'expérience globale a été jugée positive ou neutre par la plupart. Si 54 utilisateurs ont préféré la fonction de surlignage, 42 sont restés partagés sur les scores de confiance, qui ont reçu le plus de retours négatifs. 51 participants ont souhaité plus d'informations sur les erreurs potentielles, notamment les sources contradictoires.

L'analyse des réponses ouvertes a permis de faire ressortir plusieurs axes d'amélioration prioritaires :

- Sources : 11 utilisateurs ont exprimé le besoin d'avoir davantage de références scientifiques pour appuyer les réponses. Ils souhaitaient également pouvoir identifier facilement les passages précis des sources correspondant aux affirmations.
- Sources : Plus de sujets : 6 utilisateurs ont souhaité pouvoir échanger avec l'interface sur des thèmes plus variés.
- Palette de couleurs : 7 utilisateurs ont trouvé les couleurs trop vives et ont confirmé qu'il serait utile de pouvoir les personnaliser ou les désactiver.
- Historique : 7 utilisateurs aimeraient que l'application conserve leurs précédents messages.
- Tutoriel : 8 utilisateurs ont indiqué qu'un tutoriel les aiderait à mieux comprendre l'utilisation

Nous avons développé une nouvelle version de l'interface en conditions réelles, intégrant ces retours pour optimiser l'expérience utilisateur. Contrairement à la version utilisée lors de l'étude, elle utilise une méthode RAG pour enrichir les réponses, avec des articles récupérés via l'API Elsevier. Pour des contraintes techniques, seuls les résumés des articles alimentent le contexte du modèle.

3 Conception du prototype

3.1 Exigences

D'après les retours des utilisateurs, le prototype doit répondre aux exigences suivantes :

- La fonction de surlignage doit être conservée, mais laissée en option.
- Les couleurs utilisées pour le surlignage doivent être plus douces, moins vives, pour ne pas fatiguer la vue ni être trop envahissantes.
- Le score de confiance attribué aux sources doit être supprimé, car il n'a pas été jugé utile.
- Seules des sources scientifiques doivent être intégrées pour garantir crédibilité et satisfaction.
- Les utilisateurs souhaitent pouvoir identifier facilement la partie précise de la source qui soutient une affirmation, par exemple via des infobulles interactives ou des sections surlignées renvoyant directement au passage concerné.
- L'application doit élargir ses sujets pour enrichir l'expérience et toucher un public plus large.
- Le prototype doit intégrer un historique, permettant aux utilisateurs de conserver et de consulter leurs échanges précédents, pour une expérience plus personnalisée et intuitive.
- Un tutoriel doit être ajouté dans la version finale, afin d'accompagner les utilisateurs dans la prise en main de l'interface, en particulier les nouveaux venus.

3.2 Présentation générale du système

Pour répondre à ces besoins, nous présentons un prototype exploitant des LLM pour traiter un large éventail de questions à partir d'un vaste corpus académique. Notre solution combine plusieurs modules interconnectés pour extraire, analyser et intégrer les informations scientifiques pertinentes aux questions des utilisateurs. Voici les principaux éléments du système :

- Le module de recherche d'articles utilise la requête de l'utilisateur pour interroger l'API Elsevier Scopus. Il récupère les informations essentielles comme le titre, le résumé et le DOI des articles pertinents.

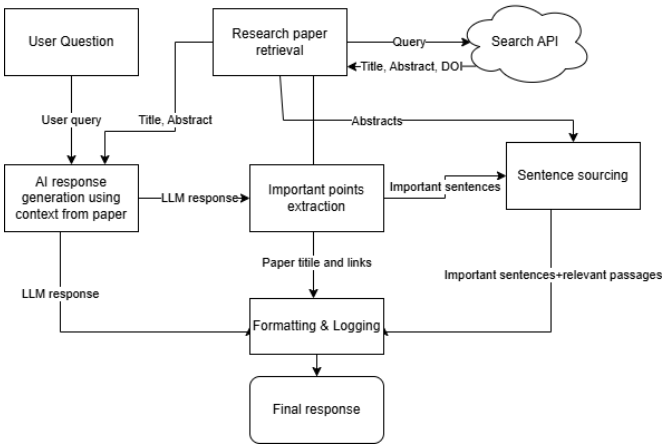

FIGURE 2 – Schéma général du système

— Le module de génération de réponse par IA se sert des informations extraites pour contextualiser et répondre précisément à la question posée.
— Le module d'extraction d'informations importantes analyse la réponse générée par le LLM afin d'en isoler les phrases clés.
— Le module de référencement des phrases associe ces phrases clés aux passages correspondants dans les résumés des articles.
— Enfin, le module de mise en forme et de journalisation regroupe la réponse de l'IA, les sources et les extraits référencés pour les afficher clairement à l'écran.

La Figure 2 présente une vue d'ensemble du système. La fonction de surlignage reste disponible si l'utilisateur choisit d'afficher les affirmations confirmées : les phrases clés sont alors surlignées en vert. Celles contredites par les résumés apparaissent en rouge et sont signalées comme potentielles désinformations. Les couleurs utilisées sont volontairement moins vives afin de réduire la fatigue visuelle. Lorsque l'utilisateur survole une zone surlignée, la phrase correspondante du résumé s'affiche pour apporter un contexte précis. Contrairement à la version initiale, si les surlignages sont désactivés, les articles restent visibles grâce à l'option « Toujours afficher les sources ». Un historique a été ajouté, permettant de consulter les questions et réponses de la conversation, avec surlignage conservé pour faciliter la relecture.

3.3 Détails de la mise en œuvre

Le système s'appuie sur plusieurs bibliothèques clés. Parmi elles, SentenceTransformer (Reimers & Gurevych, 2019) permet de générer des embeddings de phrases, FAISS (Douze *et al.*, 2025) assure une recherche efficace de similarité, et les transformers de Hugging Face sont utilisés pour charger de LLM tels que LLaMA (Touvron *et al.*, 2023), capables de produire des réponses proches du langage humain. Pour identifier les phrases clés, nous avons sélectionné les 3 plus proches de la requête utilisateur selon le modèle all-MiniLM-L6-v2. Chacune a été alignée à la phrase la plus similaire des résumés d'articles via la similarité cosinus (paraphrase-MiniLM-L6-v2). Un modèle NLI (facebook/bart-large-mnli) a ensuite déterminé si elles étaient en accord (VERT), en contradiction (ROUGE), ou neutres (sans couleur), selon l'étiquette fournie, sans seuil personnalisé. Les modèles sont exécutés sur GPU grâce à PyTorch (Paszke *et al.*, 2019), tandis que Gradio (Abid *et al.*, 2019) offre une interface conviviale facilitant l'interaction avec l'utilisateur.

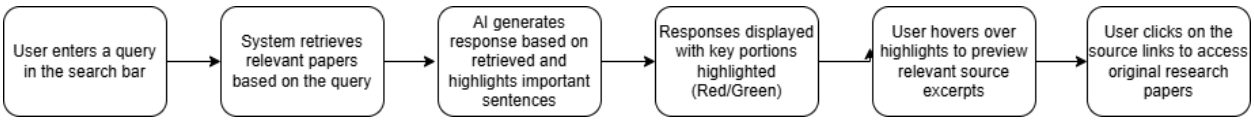

FIGURE 3 – Scénario de démonstration

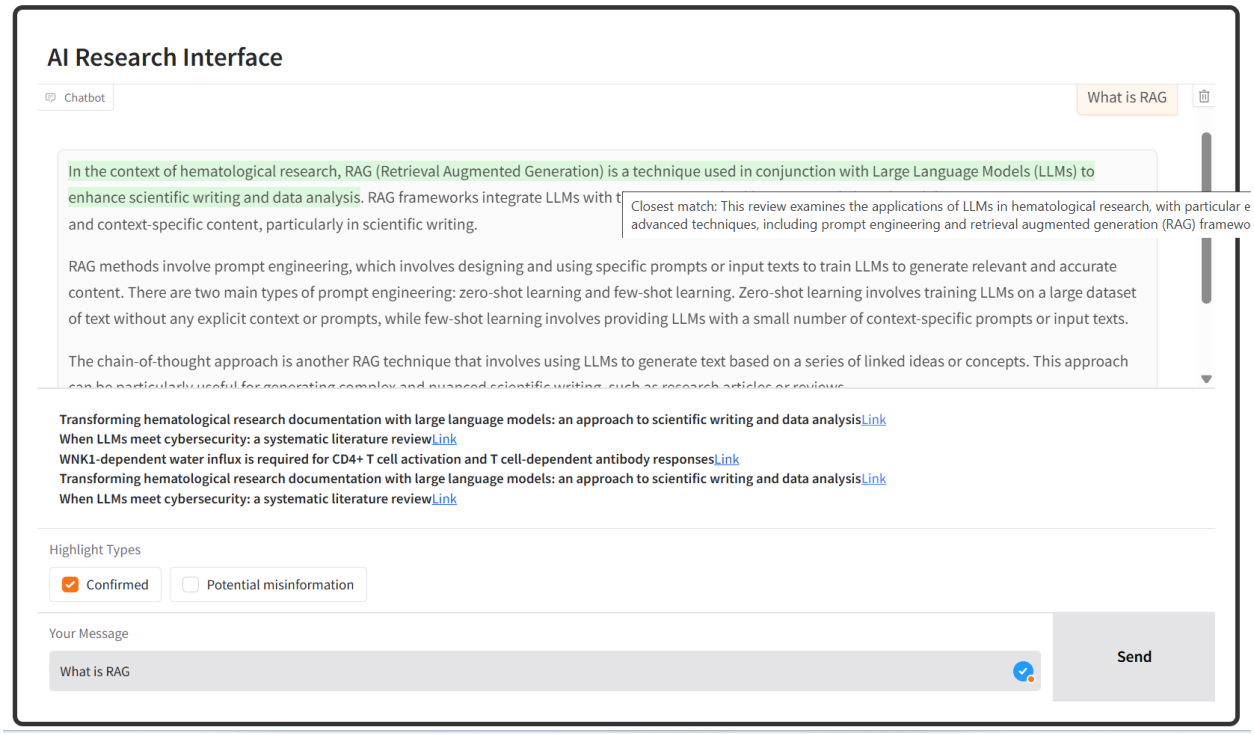

FIGURE 4 – Interface RAG finale : référencement des sources avec mise en évidence du contenu

4 Scénarios de démonstration

Ce démonstrateur cible en priorité chercheurs, doctorants et concepteurs de systèmes d'IA fiables, notamment en contexte académique où la traçabilité des sources est essentielle. L'utilisateur commence son interaction avec l'interface en saisissant une requête dans la barre de recherche. Le système récupère alors les informations pertinentes et met en évidence les passages clés pour faciliter la navigation. Les surlignages sont codés par couleur : le rouge signale des affirmations contradictoires, tandis que le vert souligne les propos appuyés par des sources fiables. La réponse de l'IA s'appuie sur les sources, garantissant une fiabilité académique. En survolant les passages surlignés, l'utilisateur peut consulter des extraits des sources, et il a la possibilité d'activer ou désactiver certaines fonctionnalités selon ses préférences. Un clic sur les liens des sources ouvre les articles complets, permettant d'explorer davantage de détails, d'activer ou non les surlignages, et de parcourir librement le contenu. Le scénario de démonstration est illustré à la Figure 3.

Optimisée pour une utilisation aussi bien sur ordinateur que sur mobile, l'interface offre une expérience fluide et agréable. Son design intuitif réduit la charge cognitive, la rendant accessible à tous les profils d'utilisateurs. Le prototype final de cette interface est présenté à la Figure 4.

5 Évaluation

Notre évaluation porte sur la conformité du prototype aux objectifs initiaux et aux attentes des utilisateurs. Bien qu'aucune étude utilisateur formelle n'ait encore été réalisée sur la version finale, les premiers retours mettent en lumière plusieurs améliorations clés visant à renforcer l'ergonomie et la confiance dans le système. Les utilisateurs ont apprécié la fonction de surlignage, tout en préférant qu'elle reste optionnelle et utilise des couleurs plus douces afin de limiter la fatigue visuelle. La suppression du score de confiance associé aux sources a également été bien reçue, soulignant l'importance de s'appuyer exclusivement sur des sources scientifiques pour garantir la crédibilité. Par ailleurs, les utilisateurs ont exprimé le besoin d'une meilleure traçabilité entre les affirmations et leurs sources, suggérant l'ajout d'infobulles interactives ou de liens directs vers les passages pertinents des textes. Nous avons élargi le champ des sujets abordés par le système et intégré une fonction d'historique pour suivre les échanges précédents, deux priorités identifiées pour améliorer l'engagement et la facilité d'utilisation. Ces observations orienteront les prochaines améliorations.

6 Conclusions

Cet article présente un prototype pour améliorer la recherche académique assistée par IA, intégrant récupération de sources fiables, génération contextuelle et attribution transparente. Basé sur l'API Elsevier Scopus, le système renforce la crédibilité des réponses et offre des outils intuitifs de vérification aux utilisateurs. La démonstration est accessible à l'adresse suivante : https://simpletext-project.com/enable/

Les retours utilisateurs ont guidé l'amélioration du système, en intégrant une option de surlignage, fonction d'historique, et meilleure couverture thématique. L'interface a été bien accueillie, surtout le surlignage, mais les avis mitigés sur les scores de confiance soulignent le besoin d'autres méthodes d'évaluation de la crédibilité. L'étude a également souligné l'importance de personnaliser les éléments visuels, comme des couleurs plus douces, pour améliorer le confort d'utilisation.

En résumé, cette recherche met en lumière l'importance d'une conception centrée sur l'utilisateur pour les systèmes d'IA en contexte académique. En permettant aux utilisateurs d'interagir de manière critique avec le contenu généré par l'IA, tout en assurant transparence et facilité d'usage, ce système pose les bases d'outils de recherche plus fiables et interactifs. Avec un développement continu, il pourrait renforcer la confiance et l'accessibilité des outils d'investigation assistés par IA.

Malgré ses fonctionnalités prometteuses, ce prototype a une couverture thématique restreinte, utilis des seuls résumés d'articles au lieu des textes complets, et des LLM encore perfectibles. Les prochaines versions viseront à élargir les sources, améliorer la précision des citations et renforcer l'accessibilité pour un public plus large. À l'avenir, nous continuerons à développer cette interface, en y intégrant un tutoriel d'accueil et de nouvelles études utilisateurs pour valider son adoption.

7 Remerciements

Ce projet bénéficie d'une subvention gouvernementale gérée par l'Agence Nationale de la Recherche via le programme «Investissements d'avenir» intégré à France 2030 (réf. ANR-19-GURE-0001).

Références

(2025). *Under evaluation*.

ABID A., ABDALLA A., ABID A., KHAN D., ALFOZAN A. & ZOU J. (2019). Gradio : Hassle-free sharing and testing of ml models in the wild.

ALI N. F., MOHTASIM M. M., MOSHARROF S. & KRISHNA T. G. (2024). Automated literature review using nlp techniques and llm-based retrieval-augmented generation. *arXiv preprint arXiv :2411.18583*.

AN H., NARECHANIA A., WALL E. & XU K. (2024). Vitality 2 : Reviewing academic literature using large language models. *arXiv preprint arXiv :2408.13450*.

ATEIA S. & KRUSCHWITZ U. (2023). Is ChatGPT a Biomedical Expert ? – Exploring the Zero-Shot Performance of Current GPT Models in Biomedical Tasks. arXiv :2306.16108 [cs], DOI : 10.48550/arXiv.2306.16108.

BELOSEVIC M. & BUSCHMEIER H. (2024). Calibrating trust and enhancing user agency in llm-based chatbots through conversational styles. *CUI@ CHI 2024 : Building Trust in CUIs–From Design to Deployment*.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

CHEN X., WANG L., WU W., TANG Q. & LIU Y. (2024). Honest ai : Fine-tuning" small" language models to say" i don't know", and reducing hallucination in rag. *arXiv preprint arXiv :2410.09699*.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DOUZE M., GUZHVA A., DENG C., JOHNSON J., SZILVASY G., MAZARÉ P.-E., LOMELI M., HOSSEINI L. & JÉGOU H. (2025). The faiss library.

ERMAKOVA L., BERTIN S., MCCOMBIE H. & KAMPS J. (2023). Overview of the clef 2023 simpletext task 3 : Simplification of scientific texts.

ERMAKOVA L., LAIMÉ V., MCCOMBIE H. & KAMPS J. (2024a). Overview of the CLEF 2024 SimpleText Task 3 : Simplify Scientific Text. In G. FAGGIOLI, N. FERRO, P. GALUŠČÁKOVÁ & A. GARCÍA SECO DE HERRERA, Éds., *Working Notes of CLEF 2024 : Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, p. 3147–3162 : CEUR-WS.org.

ERMAKOVA L., SANJUAN E., HUET S., AZARBONYAD H., DI NUNZIO G. M., VEZZANI F., D'SOUZA J. & KAMPS J. (2024b). Overview of the CLEF 2024 SimpleText track : Improving access to scientific texts for everyone. In L. GOEURIOT, G. Q. PHILIPPE MULHEM, D. SCHWAB, L. SOULIER, G. M. D. NUNZIO, P. GALUŠČÁKOVÁ, A. G. S. DE HERRERA, G. FAGGIOLI & N. FERRO, Éds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science : Springer.

GAO L., DAI Z., PASUPAT P., CHEN A., CHAGANTY A. T., FAN Y., ZHAO V. Y., LAO N., LEE H., JUAN D.-C. *et al.* (2022). Rarr : Researching and revising what language models say, using language models. *arXiv preprint arXiv :2210.08726*.

GHANEM D., ZHU A. R., KAGABO W., OSGOOD G. & SHAFIQ B. (2024). Chatgpt-4 knows its abcde but cannot cite its source. *JBJS Open Access*, **9**(3), e24.

JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y. J., MADOTTO A. & FUNG P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, **55**(12), 1–38. DOI : 10.1145/3571730.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEISER F., ECKHARDT S., LEUTHE V., KNAEBLE M., MÄDCHE A., SCHWABE G. & SUNYAEV A. (2024). Hill : A hallucination identifier for large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3613904.3642428.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., TAU YIH W., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.

LIKERT R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

PAPENMEIER A., KERN D., ENGLEBIENNE G. & SEIFERT C. (2022). It's complicated : The relationship between user trust, model accuracy and explanations in ai. *ACM Trans. Comput.-Hum. Interact.*, **29**(4). DOI : 10.1145/3495013.

PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KÖPF A., YANG E., DEVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J. & CHINTALA S. (2019). Pytorch : An imperative style, high-performance deep learning library.

PRESS O., HOCHLEHNERT A., PRABHU A., UDANDARAO V., PRESS O. & BETHGE M. (2024). Citeme : Can language models accurately cite scientific claims ?

PRIDE D., CANCELLIERI M. & KNOTH P. (2023). Core-gpt : Combining open access research and large language models for credible, trustworthy question answering.

REIMERS N. & GUREVYCH I. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* : Association for Computational Linguistics.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

THOPPILAN R., DE FREITAS D., HALL J., SHAZEER N., KULSHRESHTHA A., CHENG H.-T., JIN A., BOS T., BAKER L., DU Y. *et al.* (2022). Lamda : Language models for dialog applications. *arXiv preprint arXiv :2201.08239*.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S., BIKEL D., BLECHER L., FERRER C. C., CHEN M., CUCURULL G., ESIOBU D., FERNANDES J., FU J., FU W., FULLER B., GAO C., GOSWAMI V., GOYAL N., HARTSHORN A., HOSSEINI S., HOU R., INAN H., KARDAS M., KERKEZ V., KHABSA M., KLOUMANN I., KORENEV A., KOURA P. S., LACHAUX M.-A., LAVRIL T., LEE J., LISKOVICH D., LU Y., MAO Y., MARTINET X., MIHAYLOV T., MISHRA P., MOLYBOG I., NIE Y., POULTON A., REIZENSTEIN J., RUNGTA R., SALADI K., SCHELTEN A., SILVA R., SMITH E. M., SUBRAMANIAN R., TAN X. E., TANG B., TAYLOR R., WILLIAMS A., KUAN J. X., XU P., YAN Z., ZAROV I., ZHANG Y., FAN A., KAMBADUR M., NARANG S., RODRIGUEZ A., STOJNIC R., EDUNOV S. & SCIALOM T. (2023). Llama 2 : Open foundation and fine-tuned chat models.

WANG C., LONG Q., XIAO M., CAI X., WU C., MENG Z., WANG X. & ZHOU Y. (2024a). Biorag : A rag-llm framework for biological question reasoning. *arXiv preprint arXiv :2408.01107*.

WANG T., HE J. & XIONG C. (2024b). Ragviz : Diagnose and visualize retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 320–327.

YAN S.-Q., GU J.-C., ZHU Y. & LING Z.-H. (2024). Corrective retrieval augmented generation.

YIN C., WEI E., ZHANG Z. & ZHAN Z. (2025). Paperhelper : Knowledge-based llm qa paper reading assistant. *arXiv preprint arXiv :2502.14271*.

ZHOU K., HWANG J. D., REN X., DZIRI N., JURAFSKY D. & SAP M. (2024). Rel-a.i. : An interaction-centered approach to measuring human-lm reliance.

ZUCCON G., KOOPMAN B. & SHAIK R. (2023). ChatGPT Hallucinates when Attributing Answers. arXiv :2309.09401 [cs], DOI : 10.48550/arXiv.2309.09401.