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

Vers des RAGs intégrant véracité, subjectivité et explicabilité

Alae Bouchiba[1] Adrian-Gabriel Chifu[1] Sébastien Fournier[1] Lorraine Goeuriot[2]
Philippe Mulhem[2]
(1) Aix Marseille Univ, CNRS, LIS, Marseille, France
(2) Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP[*], LIG, 38000 Grenoble, France
(1) {prénom}.{nom}@lis-lab.fr, (2) {prénom}.{nom}@univ-grenoble-alpes.fr

RÉSUMÉ

Cet article introduit X-RAG-VS , un cadre pour intégrer véracité , subjectivité et explicabilité dans les systèmes RAG , en réponse aux besoins éducatifs. À travers des cas d'usage et l'analyse de modèles existants , nous montrons que ces dimensions restent insuffisamment prises en compte. Nous proposons une approche unifiée pour des réponses plus fiables , nuancées et explicables.

ABSTRACT

Towards RAGs Integrating Veracity, Subjectivity, and Explainability

We introduce X-RAG-VS , a framework aiming to integrate veracity , subjectivity , and explainability within retrieval-augmented generation systems for educational use. Based on concrete use cases and an analysis of current large language models , we show that these critical dimensions are often addressed separately or insufficiently. We propose a unified approach to support more reliable , nuanced, and transparent AI-generated content.

MOTS-CLÉS : RAG , éducation , véracité, subjectivité , pensée critique , explicabilité.

KEYWORDS: RAG , education , veracity , subjectivity , critical thinking , explainability.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

1 Introduction

Cet article propose l'étude des systèmes de génération augmentée par récupération (RAG) (Lewis *et al.*, 2020) adaptés aux besoins éducatifs , ayant pour objectif de favoriser l'esprit critique , la transparence et la pluralité des points de vue. Nous articulons nos travaux au travers des trois dimensions clés suivantes : **véracité**, **subjectivité** et **explicabilité**. Nous nous concentrons sur ces trois dimensions, nécessaires pour garantir des réponses à la fois fiables , nuancées et accessibles.

Nous proposons ainsi les fondements d'un cadre méthodologique, que nous appelons X-RAG-VS , visant à explorer comment ces dimensions , souvent traitées séparément dans la littérature , peuvent être combinées de manière cohérente pour répondre à des enjeux pédagogiques concrets : aider les apprenants à juger de la fiabilité d'une information, à confronter des opinions divergentes et à saisir les mécanismes de raisonnement d'un système d'intelligence artificielle (IA).

∗. Institute of Engineering Univ. Grenoble Alpes

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 149–155.

Cette contribution s'inscrit en tant que réflexion structurée sur les conditions d'un usage responsable de l'IA générative en contexte éducatif. Elle se positionne comme un travail exploratoire , à la croisée entre cadre théorique et mise en perspective empirique , destiné à nourrir les discussions sur la formation à l'esprit critique à l'ère des grands modèles de langage.

L'analyse qui suit est décomposée en 3 étapes : tout d'abord nous allons décrire un exemple lié au domaine de l'éducation, pour lequel les 3 dimensions sont nécessaires conjointement. Tout comme dans (Chen *et al.*, 2023), on peut se poser la question de savoir dans quelle direction faire porter le travail de recherche : étudier les prompts ou bien modifier les modèles LLMs. Dans un second temps, nous allons montrer que ces dimensions sont grandement absentes dans les réponses générées par deux modèles de chat connus , au travers de prompts simples. Nous allons ensuite explorer dans quelle mesure les approches de la littérature sont capables d'intégrer tous ces éléments. Nous proposerons ensuite un plan de travail pour proposer de dépasser les limites existantes.

2 Cas d'usage et axes d'analyse

Afin d'analyser les besoins et l'existant pour nos trois dimensions d'étude , nous définissons dans la suite un cadre d'analyse basé sur des cas d'usage.

Les cas d'usage sont , pour nous , définis : i) par des étudiants ou des écoliers , ii) dans le cadre de la recherche d'un panorama d'avis sur un sujet de société. Ces cas d'usage se situent plus globalement dans l'"éducation à la pensée critique" : cette dernière repose sur des arguments et des faits et sur leur mise en perspective argumentée, qui eux-mêmes reposent donc sur des éléments pour lesquels la véracité est caractérisée (est-ce que les arguments sont vrais) , pour lesquels la subjectivité/objectivité est estimée (est-ce que les éléments sont objectifs ou subjectifs) , et enfin pour lesquels des explications sur les raisons de la présentation des éléments sont fournies (pourquoi ces arguments sont présentés).

Plus précisément :

— **la véracité** est entendue comme la capacité à produire des informations exactes (Hwang *et al.*, 2024) et fondées sur des sources fiables (Zhou *et al.*, 2024b) ;

— **la subjectivité** , qui renvoie à la prise en compte de points de vue divergents (Wan & McAuley, 2016), d'opinions ou de jugements de valeur (Balayn & Bozzon, 2019) ;

— **l'explicabilité** , c'est-à-dire la faculté pour le système d'expliquer de manière intelligible les choix réalisés, les sources mobilisées ou les raisonnements suivis (Arrieta *et al.*, 2020).

Ces dimensions sont aujourd'hui considérées comme essentielles pour concevoir des systèmes d'IA plus transparents, nuancés et pédagogiquement adaptés (Maity & Deroy, 2024) et (Karran *et al.*, 2024).

Le succès des LLMs pour nos cas d'usage pourraient grandement bénéficier de la prise en compte **conjointement** de ces trois dimensions : les textes générés seraient alors par exemple capables d'expliquer pourquoi (explicabilité) un point de vue qui est supporté par des informations vraies (véracité) mais supportées par des sources d'information subjectives (subjectivité) est choisi pour être présenté dans l'argumentaire de la réponse, car ce point de vue est l'un des éléments qui permet d'expliquer la complexité d'un sujet.

3 Utilisation de prompts pour les trois dimensions explorées

Cette section a pour objectif d'estimer dans quelle mesure certains LLMs de chat intègrent par défaut certains éléments des dimensions de véracité, de subjectivité et d'explicabilité.

L'idée d'explorer comment les prompts seraient capables d'intégrer la prise en compte de ces trois dimensions vient immédiatement à l'esprit, car il a été montré (Zhou *et al.*, 2024a) que le prompt-tuning a un impact important sur la qualité des réponses de ces modèles sans nécessiter d'apprentissage. Afin d'évaluer empiriquement cette hypothèse , cette étude propose une expérimentation originale utilisant deux LLMs dédiés aux *chat* , ChatGPT 4o en mode "recherche web" et Gemini 2.5 Pro. Bien qu'ils ne soient pas des RAG , ils sont capables de sourcer (i.e., indiquer des références) leur réponse , ce qui nous permet d'étudier leur comportement par rapport à nos dimensions d'intérêt. Notre étude vise à analyser , pour un contexte d'utilisation spécifique en se basant sur des prompts simples , si et comment ces systèmes sont capables d'intégrer ces trois dimensions de manière satisfaisante.

Les deux requêtes soumises dans cette étude correspondent au niveau 1 de la taxonomie TELeR (Santu & Feng, 2023) , le choix de celles-ci s'est basé sur le choix de l'utilisateur , pour délibérément imiter ce qu'un collégien ou un lycéen les formuleraient. Nous testons la capacité du système à explorer les dimensions naturellement sans pour autant complexifier la requête . Le choix des deux requêtes était : « Faut-il interdire les devoirs à la maison ?» et «Faut-il raccourcir les vacances d'été ?».

Dans cette perspective, nous avons analysé les réponses générées par ces deux grands modèles de langues.

Les deux modèles de langues utilisés présentent une analogie structurée dans leur manière d'aborder le sujet ; ils adoptent la même posture argumentative , exposant de façon équilibrée les points favorables et défavorables à la question posée.

L'analyse détaillée exhibe cependant les éléments suivants :

— La symétrie des arguments dans la présentation des idées donne l'illusion d'une neutralité, mais masque en réalité une absence d'engagement critique et ne laisse pas la place à une véritable pluralité de points de vue vécus , ni à l'expression des différentes positions incarnées.

— Le choix des références mobilisées, souvent issues de blogs ou de sites web non académiques, renforce cette impression de surface argumentative : les contenus sont illustratifs , mais ne s'appuient pas sur des sources clairement identifiées comme fiables.

— Le raisonnement sous-jacent et le processus de sélection des sources sont peu ou pas explicités dans les réponses générées. Cette opacité rend difficile l'identification des parties du texte sur lesquelles le modèle s'appuie réellement pour construire son argumentation , et empêche de comprendre comment il réagit, structure sa réflexion, ou mobilise des connaissances face à la question posée. On a donc aucune information d'explication fournie par ces modèles.

Aucun des deux modèles testés ne parvient à intégrer de manière cohérente et satisfaisante l'ensemble des dimensions ciblées. Si certains éléments , comme la structuration explicative ou la présence d'arguments opposés , sont prometteurs , ils restent insuffisants en l'absence de vérification documentaire rigoureuse et de véritable diversité argumentative.

On pourrait arguer que des prompts plus complexes seraient peut-être à même d'améliorer la prise en compte de nos dimensions d'analyse. IL a cependant été montré récemment (Zou *et al.*, 2023).

la forte sensibilité des LLMs à la formulation des prompts , ainsi que les risques d'hallucinations persistantes dans des contextes sensibles (Dahl *et al.*, 2024). Ces observations confirment selon nous que le prompt-tuning [1] seul ne constitue pas une solution robuste pour les enjeux complexes que nous visons.

Ces résultats confirment alors , dans notre cas d'usage , que les trois dimensions identifiées comme essentielles à la pensée critique sont très inégalement prises en compte dans les modèles actuels . Cela rejoint les constats établis dans la littérature scientifique , que nous présentons dans la section suivante.

4 Les dimensions d'analyse dans la littérature

Dans cette section , nous examinons dans quelle mesure la littérature scientifique actuelle propose des réponses à notre problématique , en explorant des travaux portant explicitement sur la véracité , la subjectivité ou l'explicabilité dans des systèmes similaires.

4.1 Véracité

La véracité ne se réduit pas à la cohérence interne des réponses générées : elle implique la capacité du système à produire une information exacte , fondée sur des sources identifiables. Cette exigence est d'autant plus cruciale dans un cadre éducatif , où le développement de la pensée critique suppose que l'apprenant puisse évaluer la qualité des connaissances mobilisées. tout comme l'étude VeraCT (Chen *et al.*, 2024) qui affirme cette nécessité et propose une approche de vérification des faits par récupération et génération , dans laquelle chaque affirmation est confrontée à des sources externes crédibles , accompagnées d'un raisonnement structuré. Dans une perspective plus large , la synthèse (Zhao *et al.*, 2024) met en évidence les différentes stratégies de vérification de revendications intégrées aux LLMs , soulignant que la véracité constitue un pilier fondamental pour juger de la robustesse d'un système. Ce principe reste le même pour d'autres travaux, tout en évaluant la fiabilité des sources web à partir de la qualité factuelle des informations qu'elles véhiculent (Dong *et al.*, 2015) . Ces approches convergent vers une même conclusion : pour qu'un système RAG contribue réellement à l'apprentissage , il doit offrir un accès structuré à une information vérifiée , bien que traçable.

4.2 Subjectivité

Dans un contexte éducatif et de formation à l'esprit critique , il est essentiel que les systèmes soient capables de restituer non seulement des faits vérifiés , mais aussi la diversité des points de vue présents sur un sujet. La capacité à exposer des opinions divergentes est particulièrement cruciale dans les situations de débat ou d'analyse de controverses , où il n'existe pas une vérité unique mais plusieurs interprétations légitimes. Pourtant, cette dimension de subjectivité reste largement absente des architectures RAG actuelles. Des travaux récents , comme pour (Chen *et al.*, 2024) , mettent en évidence la nécessité d'intégrer des perspectives multiples pour enrichir la génération de contenu. De même , l'approche Vendi-RAG proposée (Rezaei & Dieng, 2025) introduit un mécanisme d'optimisation adaptative entre diversité et qualité , permettant de générer des réponses plus nuancées et représentatives de la pluralité des opinions. Dans le cadre éducatif , intégrer la subjectivité aux côtés de la véracité et de l'explicabilité est indispensable pour développer des outils qui ne se contentent

1. Le fait de modifier les prompts pour affiner la réponse, de manière experte.

pas de transmettre une information correcte , mais qui permettent aux apprenants de comprendre la complexité des débats , de comparer les arguments , et de forger leur propre jugement critique.

4.3 Explicabilité

l'explicabilité ne se limite pas à exposer les documents utilisés, mais implique la capacité du système RAG à articuler de manière intelligible le raisonnement qui relie ces sources à la réponse générée , cette forme d'explicabilité est particulièrement précieuse dans le cadre éducatif , où la compréhension du raisonnement importe autant que le résultat produit. Dans ce sens , (Xu *et al.*, 2023) proposent une approche de RAG interprétable , dans laquelle chaque segment de réponse est justifié par un élément explicite du document source , renforçant ainsi la traçabilité et la vérifiabilité de l'information générée. De manière complémentaire , (Tian *et al.*, 2023) introduisent un pipeline de génération structuré où les sources récupérées sont non seulement affichées , mais réorganisées dans un graphe d'argumentation destiné à expliciter la logique suivie par le modèle. Ces efforts montrent que l'explicabilité dans RAG ne peut être atteinte uniquement par la transparence des sources , mais nécessite également une mise en forme narrative et rationnelle de l'information.

D'autres travaux visent également à combiner plusieurs de nos dimensions clés , comme pour (Abolghasemi *et al.*, 2024) qui s'intéressent à la manière dont les biais d'attribution influencent à la fois la fiabilité des sources mobilisées et la pluralité des perspectives exposées, articulant ainsi véracité et subjectivité. Par ailleurs , (Rezaei & Dieng, 2025) propose un cadre adaptatif qui équilibre diversité argumentative et cohérence des réponses générées , contribuant à une meilleure structuration du contenu et à une forme d'explicabilité implicite , sans pour autant rassembler les trois.

- Notre analyse ci-dessus révèle ainsi des propositions qui sont fragmentées , car chacune des dimensions qui nous intéressent sont traitées de manière indépendante , ou dépendante mais incomplète. Nous défendons le point de vue qu'une approche unifiée serait plus à même d'être capable d'articuler conjointement les dimensions de véracité , d'explicabilité et de subjectivité dans une même architecture.

- Il convient de noter que bien que les trois dimensions de véracité , subjectivité et explicabilité puissent présenter des interdépendances. Par exemple , l'explicabilité peut contribuer à éclairer la fiabilité des sources (véracité) ou la diversité des perspectives présentées (subjectivité). Cette analyse les considère comme des dimensions distinctes et autonomes. Cette approche méthodologique permet d'identifier plus précisément les lacunes spécifiques à chaque dimension dans les systèmes actuels et de concevoir des solutions ciblées pour chacune d'entre elles , avant d'envisager leur intégration cohérente dans un cadre unifié. De plus , cette distinction analytique facilite l'évaluation comparative des différentes approches de la littérature et permet de mieux cerner les contributions spécifiques de chaque travail par rapport aux trois dimensions considérées. C'est précisément cette analyse différenciée qui révèle la nécessité d'un cadre comme X-RAG-VS , capable d'orchestrer ces dimensions de manière synergique plutôt que de les traiter comme des éléments isolés ou faiblement articulés.

5 Vers une intégration de la véracité, la subjectivité et l'explicabilité

Il apparaît , sans pour autant être des preuves définitives , d'après les quelques expérimentations menées , que ni la simple ingénierie de prompts , ni les approches existantes dans la littérature ne

permettent , à ce jour , de prendre en compte de manière intégrée les trois dimensions qui nous intéressent.

L'approche X-RAG-VS que nous proposons repose sur deux axes complémentaires. D'une part , elle intègre les dimensions de véracité , subjectivité et explicabilité directement dans le processus de génération , en agissant à un niveau sémantique fin. D'autre part , elle s'inspire de mécanismes tels que SELF-RAG (Liu, 2023) , qui déclenchent dynamiquement la recherche documentaire , adaptés ici à des finalités éducatives.

Une telle prend tout son sens dans des contextes d'apprentissage concrets. Par exemple , lorsqu'un élève cherche à se forger une opinion sur une question de société , la génération ne se limite pas à structurer une réponse : elle articule des informations vérifiables , présente une diversité de points de vue et explicite le raisonnement suivi. Le système devient alors un véritable support au développement de la pensée critique , en phase avec les cas d'usage identifiés.

Remerciements

Cette recherche a été financée en partie par l'Agence nationale de la recherche (ANR) au titre du projet GUIDANCE , ANR-23-IAS1-0003.

Références

ABOLGHASEMI A., AZZOPARDI L., HASHEMI S. H. *et al.* (2024). Evaluation of attribution bias in retrieval-augmented large language models. *arXiv preprint arXiv :2410.12380*.

ARRIETA A. B., DÍAZ-RODRÍGUEZ N., SER J. D., BENNETOT A., TABIK S., BARBADO A., GARCIA S., GIL-LOPEZ S., MOLINA D., BENJAMINS R., CHATILA R. & HERRERA F. (2020). Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, **58**, 82–115. arXiv :1910.10045.

BALAYN A. & BOZZON A. (2019). Designing evaluations of machine learning models for subjective inference : The case of sentence toxicity. arXiv preprint arXiv :1911.02471.

CHEN B., YI F. & VARRÓ D. (2023). Prompting or fine-tuning ? a comparative study of large language models for taxonomy construction. arXiv :2309.01715.

CHEN T., SORENSEN J., ZIEMS C. *et al.* (2024). Retrieval-augmented generation with diverse perspectives. *arXiv preprint arXiv :2409.18110*.

DAHL M., MAGESH V., SUZGUN M. & HO D. E. (2024). Large legal fictions : Profiling legal hallucinations in large language models. arXiv preprint arXiv :2401.01301.

DONG X. L., GABRILOVICH E., MURPHY K., DANG V., HORN W., LUGARESI C., SUN S. & ZHANG W. (2015). Knowledge-based trust : estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, **8**(9), 938–949. DOI : 10.14778/2777598.2777603.

HWANG S., BAEK J., PARK J. & KANG J. (2024). Retrieval-augmented generation with estimation of source reliability. https://arxiv.org/abs/2410.22954. arXiv :2410.22954.

KARRAN T., HALL M. & NUNAN T. (2024). Multi-stakeholder perspective on responsible artificial intelligence and acceptability in education. *arXiv preprint arXiv :2402.15027*.

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. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. NeurIPS 2020.

LIU Z. E. A. (2023). Self-rag : Retrieval-augmented generation via self-retrieval. https://arxiv.org/abs/2310.11511. arXiv preprint arXiv :2310.11511v1.

MAITY D. & DEROY Y. (2024). Human-centric explainable ai in education. *arXiv preprint arXiv :2410.19822*.

REZAEI M. R. & DIENG A. B. (2025). Vendi-rag : Adaptively trading-off diversity and quality significantly improves retrieval augmented generation with llms. *arXiv preprint arXiv :2502.11228*.

SANTU S. K. K. & FENG D. (2023). Teler : A general taxonomy of llm prompts for benchmarking complex tasks. *arXiv preprint arXiv :2305.11430*.

TIAN Y., YE D., LIN Y., LIU Z. & SUN M. (2023). Explanation graph generation via pre-trained language models. *arXiv preprint arXiv :2305.14277*.

WAN M. & MCAULEY J. (2016). Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, p. 489–498 : IEEE. DOI : 10.1109/ICDM.2016.0060.

XU C., YU M., GAO J., GAO Z., LIN J. & CALLAN J. (2023). Towards interpretable retrieval-augmented generation : A case study on explainable qa. *arXiv preprint arXiv :2310.03667*.

ZHAO L., LIU X. & LIU Q. (2024). Claim verification in the age of large language models : A survey. *arXiv preprint arXiv :2408.14317*.

ZHOU X., BEHROOZ M., DEHGHANI M. & REN X. (2024a). The prompt report : A systematic survey of prompting techniques. arXiv preprint arXiv :2406.06608v2.

ZHOU Y., LIU Y., LI X., JIN J., QIAN H., LIU Z., LI C., DOU Z., HO T.-Y. & YU P. S. (2024b). Trustworthiness in retrieval-augmented generation systems : A survey. arXiv :2409.10102.

ZOU A., WANG Z., CARLINI N., NASR M., KOLTER J. Z. & FREDRIKSON M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv :2307.15043*.