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

Exploration du RAG pour la génération de réponses à des questions en contexte éducatif: étude sur les données SCIQ

Sarah Nouali Ismail Badache Patrice Bellot
Aix-Marseille University, Université de Toulon, CNRS, LIS, Marseille, France
{sarah.nouali, ismail.badache, patrice.bellot}@lis-lab.fr

RÉSUMÉ

Les systèmes basés sur le RAG (Retrieval-Augmented Generation) sont des systèmes qui optimisent la puissance des grands modèles de langue (LLM, en anglais, *Large Language Models*) avec une recherche d'information (RI) à partir de sources de connaissances externes, sans avoir besoin de réentraîner le modèle. Ce type d'approche est connu pour améliorer les réponses du LLM, en particulier pour répondre à des questions spécifiques à un domaine, et réduire le phénomène d'hallucination constaté avec ces derniers. Dans cet article, nous explorons l'application d'un tel système dans un contexte pédagogique, en utilisant le jeu de données SCIQ (*SCIence Questions*), un ensemble de questions scientifiques à choix multiples de niveau scolaire, qui nous permet d'évaluer la capacité des modèles à fournir des réponses précises, pédagogiques et vérifiables. Nous évaluons les performances du système par rapport à un modèle génératif standard (*Llama3 8b* et *Mistral 7b*) de réponse aux questions et analysons ses forces et ses limites dans un contexte éducatif. La performance la plus élevée en termes de précision a été enregistrée avec l'approche basée sur le RAG (*rag-llama*), qui a permis d'atteindre une précision globalement supérieure par rapport aux autres approches testées.

ABSTRACT

Exploring RAG for educational question answering : A study on the SCIQ dataset

Retrieval-Augmented Generation (RAG) based systems are systems that optimize the power of large language models (LLMs) with information retrieval from external knowledge sources, without the need to re-train the model. This type of approach is known to improve LLM responses, particularly when answering domain-specific questions, and reduce the hallucination phenomenon seen with the latter. In this article, we explore the application of such a system in a pedagogical context, using the SCIQ dataset, set of grade-level multiple-choice scientific questions, which enables us to assess the models' ability to provide accurate, pedagogical and verifiable answers. We evaluate the system's performance against a standard generative question answering model (LLM) and analyze its strengths and limitations in an educational context. The highest performance in terms of accuracy was recorded with the RAG-based approach (*rag-llama*), which achieved an overall higher accuracy than the other approaches tested.

MOTS-CLÉS : système question-réponse, grands modèles de langue, RAG, éducation, SCIQ.

KEYWORDS: question-answer system, large language models, RAG, education, SCIQ.

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

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 29–41.

1 Introduction

De nos jours, l'intelligence artificielle (IA) est de plus en plus présente dans de nombreux domaines, et l'éducation n'échappe pas à cette tendance. L'IA est désormais intégrée dans les environnements éducatifs à travers des applications variées : pour l'aide aux élèves via des systèmes de tutorat intelligent, apprentissage ludique (*game-based learning*) ou encore évaluation formative automatisée, pour apporter un soutien aux enseignants avec des outils de détection de plagiat, de curation intelligente de ressources pédagogiques ou d'évaluation sommative automatisée, ou pour fournir une assistance aux institutions et établissements scolaires grâce à des systèmes de gestion des admissions, de planification des cours et des emplois du temps, ou encore de surveillance à distance des examens (Holmes & Tuomi, 2022). Parmi ces usages, les systèmes de tutorat intelligent (STI) comptent parmi les applications d'IA les plus répandues et les mieux financées dans le domaine de l'éducation. Ils proposent des tutoriels informatisés, étape par étape, adaptés à chaque élève, principalement dans des disciplines structurées comme les mathématiques (Holmes & Tuomi, 2022).

Dans ce contexte, les systèmes de réponse automatique aux questions (QA) apparaissent comme particulièrement prometteurs pour accompagner l'apprentissage des élèves et fournir des explications à la demande. Toutefois, les modèles de langage de grande taille (LLMs), bien que puissants, génèrent leurs réponses à partir de connaissances apprises lors de l'entraînement. Cela peut entraîner des réponses erronées, obsolètes, ou encore des phénomènes d'hallucination — une faiblesse bien connue de ces systèmes. Pour pallier ces limites, les systèmes de type génération augmentée par la recherche d'information *Retrieval-Augmented Generation* (RAG) (Lewis *et al.*, 2020) ont été proposés. Ils combinent les capacités génératives des LLMs avec un module de recherche documentaire, permettant de produire des réponses fondées sur des sources externes fiables et vérifiables, telles que des articles scientifiques ou des manuels scolaires. Dans un contexte éducatif, cette capacité à s'appuyer sur des sources factuelles est essentielle. Elle garantit des réponses plus fiables, alignées avec les objectifs pédagogiques définis par les programmes scolaires. De plus, les systèmes RAG offrent une transparence accrue : les documents utilisés pour formuler la réponse peuvent être consultés, ce qui permet aussi bien aux élèves qu'aux enseignants de comprendre le raisonnement sous-jacent. Enfin, ces systèmes peuvent être adaptés à des corpus spécifiques, permettant un alignement précis avec des référentiels pédagogiques ou des supports institutionnels. Ainsi, les systèmes RAG représentent une alternative plus fiable, interprétable et potentiellement mieux adaptée aux exigences de l'enseignement que les modèles génératifs classiques.

Ce travail a pour objectif d'explorer de manière approfondie le potentiel de cette approche dans un contexte éducatif, en s'appuyant sur un ensemble de questions de recherche structurantes qui guident l'analyse et la réflexion.

— **Q1 :** Comment intégrer le RAG dans un environnement de tutorat intelligent, et quelle est son efficacité pour ce type de tâche ?

— **Q2 :** Comment l'utilisation du RAG peut-elle renforcer la confiance des utilisateurs envers les réponses du système dans un contexte éducatif où la véracité est primordiale ?

— **Q3 :** Est-ce que les systèmes basés sur le RAG permettent réellement de fournir des sources ou des supports justifiant les réponses générées ?

— **Q4 :** Les réponses générées par un système RAG sont-elles suffisamment explicites et claires pour favoriser la compréhension des apprenants ?

2 IA et RAG en éducation : une vue d'ensemble

La présente section propose une vue d'ensemble des apports de l'intelligence artificielle (IA) en éducation, en articulant deux axes complémentaires. Le premier examine les usages généraux de l'IA dans le domaine éducatif, en s'appuyant sur les évolutions récentes des technologies basées sur les modèles de langage. Le second s'intéresse plus spécifiquement aux systèmes dits RAG, conçus pour pallier certaines limites des LLMs, en particulier le phénomène d'hallucination, en intégrant des mécanismes de récupération d'informations provenant de sources externes fiables.

2.1 l'IA et l'éducation

Nous pouvons classer les solutions d'IA appliquées à l'éducation en trois catégories : celles destinées à aider les élèves dans leur apprentissage, celles conçues pour accompagner les enseignants, et enfin celles dédiées au soutien des établissements et institutions scolaires (Holmes & Tuomi, 2022). Dans la catégories des applications pour élèves, nous retrouvons plusieurs types de solutions : les systèmes de tutorat intelligent (Ward *et al.*, 2013; Dolenc *et al.*, 2015; Lane *et al.*, 2013), les environnements d'apprentissage exploratoires, les simulations assistées par l'IA comme les jeux éducatifs numériques (McLaren *et al.*, 2017; Parong *et al.*, 2017; Mayer, 2019), les agents conversationnels (Paschoal *et al.*, 2018; Sreelakshmi *et al.*, 2019; Lin, 2019; Deveci Topal *et al.*, 2021), les systèmes d'évaluation automatisée (*Automated assessment and feedback*) (Lee *et al.*, 2019; Sung *et al.*, 2021; Maestrales *et al.*, 2021), les outils de rédaction automatique comme les générateurs de textes [1] et les correcteurs grammaticaux automatisés [2], ou encore les applications assistées par l'IA tel que les outils de traduction [3] et de résolutions de problèmes mathématiques [4]. L'IA joue également un rôle important dans le soutien aux élèves en situation de handicap, grâce à des applications de diagnostic (dyslexie, TDAH, dysgraphie) (Barua *et al.*, 2022) et à des outils d'assistance (sous-titrage automatique, synthèse vocale, etc.). Pour les enseignants, nous pouvons trouver les détecteurs de plagiat [5], les plateformes de curation intelligente de contenus [6], les systèmes de monitoring en classe (Bosch & D'Mello, 2021), et les assistants à l'évaluation et à la correction. Enfin, du côté des institutions, nous allons voir des solutions d'IA pour optimiser la planification des cours (Kitto *et al.*, 2020), la détection des élèves à risque (Del Bonifro *et al.*, 2020), la gestion des admissions (Marcinkowski *et al.*, 2020), ou encore la surveillance automatisée des examens (e-proctoring) (Nigam *et al.*, 2021).

Avec l'apparition de l'IA générative, un nouveau paradigme est né, en particulier pour les systèmes de tutorat intelligents. En effet, grâce aux grands modèles de langue, il est maintenant possible de générer des contenus éducatifs dynamiques et pertinents en fonction du contexte (Maity & Deroy, 2024). Une des applications de l'IA générative concerne les systèmes de dialogue interactifs. Ces modèles permettent des échanges plus naturels et engageants entre l'apprenant et le système pour expliquer et répondre à ses questions. Malgré les avantages que l'intégration de l'IA générative apporte à l'éducation, en particulier dans les systèmes tutoriels intelligents, cette pratique soulève plusieurs défis majeurs. En effet, un des problèmes connus des LLMs est la possibilité de production d'informations erronées ou inappropriées, ainsi que l'impact de biais (stéréotypes, surreprésentation

1. https ://chatgpt.com/g/g-OolQ7FMzJ-ai-text-generator-gpt
2. https ://www.grammarly.com/
3. https ://www.deepl.com/translator
4. https ://photomath.com/
5. https ://plagiarismcheckerx.com, https ://www.turnitin.com/
6. https ://www.x5gon.org/

de certains points de vue, influence de la localisation des informations dans les documents...) durant l'apprentissage. Or, dans un contexte éducatif, pouvoir garantir la pertinence pédagogique du contenu généré est essentiel (Maity & Deroy, 2024).

Face aux défis soulevés par l'usage de l'IA générative dans les environnements éducatifs, il est nécessaire de concevoir des approches ciblées et fiables. C'est dans cette optique que nous proposons un système de question-réponse capable de répondre à des questions spécifiques en exploitant des supports privilégiés pour appuyer les réponses données.

2.2 Les systèmes RAG et l'éducation

Une des solutions qui a émergé pour atténuer ou pallier les failles des LLMs, en particulier le phénomène d'hallucination, est le *Retrieval Augmented Generation* (RAG) (Swacha & Gracel, 2025). Le RAG améliore les performances des LLMs en combinant une récupération d'informations pertinentes dans une base de données (ou base documentaire de référence) avec la génération de réponses en langage naturel, offrant ainsi des réponses plus précises et contextuellement adaptées.

Plusieurs applications existent dans la littérature qui intègrent le RAG afin d'interroger une source de connaissance externe : en domaine juridique (Wiratunga *et al.*, 2024; Cui *et al.*, 2024), pour les sciences de la santé et des sciences biomédicales (Li *et al.*, 2023; Lála *et al.*, 2023), la finance (Habib *et al.*, 2024), l'informatique (Dean *et al.*, 2023) ou encore pour plusieurs domaines (Forootani *et al.*, 2025). Malgré que ces applications ne sont pas spécifiquement faitent pour l'apprentissage, ce type de systèmes peut-être utilisé dans un contexte éducatif. Nous retrouvons bien évidemment plusieurs approches destinées à l'apprentissage (Jiang *et al.*, 2024; Abraham *et al.*, 2024; Levonian *et al.*, 2023; Al Ghadban *et al.*, 2023).

3 Approche et Méthodologie

Cette section présente de manière détaillée les fondements méthodologiques de notre étude, en articulant d'abord la description du jeu de données Science Questions (*SCIQ*) utilisé comme collection de tests, puis en exposant les différents systèmes et configurations expérimentales testés et expérimentés pour répondre aux problématiques posées, avant de détailler enfin la stratégie d'évaluation mise en œuvre pour mesurer rigoureusement la performance et la pertinence des approches proposées.

3.1 Jeu de données Science Questions "SCIQ"

Nous avons choisi d'utiliser l'ensemble de données SCIQ (SCIence Questions)[7] (Welbl *et al.*, 2017) pour évaluer notre système RAG dans un contexte éducatif. SCIQ a été obtenu par *crowdsourcing* et a été conçu pour l'entraînement et l'évaluation de modèles de question-réponse (Welbl *et al.*, 2017). Il se compose de plus de 13.000 questions à choix multiples, découpé en trois sous-ensembles (11.700 questions dans l'ensemble d'entraînement et 1000 questions dans chaque ensemble de validation et de test) couvrent un niveau allant de l'école primaire aux cours d'introduction à l'université, telles que la biologie, la physique, la chimie et les sciences de la Terre (Welbl *et al.*, 2017). Chaque élément

7. https ://huggingface.co/datasets/allenai/sciq

de cet ensemble contient une question, quatre réponses possibles, parmi lesquelles une seule est correcte, la majorité des questions sont accompagnées de paragraphes supplémentaires et d'éléments d'information à l'appui des bonnes réponses (Yu *et al.*, 2024). Ce jeu de données est en anglais.

Question	*What type of organism is commonly used in preparation of foods such as cheese and yogurt ?*
Réponses	**A : mesophilic organisms** B : protozoa C : gymnosperms D : viruses
Support / Explication	*Mesophiles grow best in moderate temperature, typically between 25 C and 40 C (77 F and 104 F). Mesophiles are often found living in or on the bodies of humans or other animals. The optimal growth temperature of many pathogenic mesophiles is 37 C (98 F), the normal human body temperature. Mesophilic organisms have important uses in food preparation, including cheese, yogurt, beer and wine.*

TABLE 1 – Un exemple d'entrée de SCIQ, constitué d'une question, de quatre réponses (la bonne réponse est en gras), ainsi que du passage justifiant la bonne réponse.

Notre choix s'est porté sur ce jeu de données pour les raisons suivantes : d'un côté, il constitue un ensemble standard, reconnu pour l'évaluation de modèles de type Question-Réponse (*Question-Answering*, QA) dans un cadre éducatif (Liu *et al.*, 2024). En outre, les questions sont formulées de manière simple mais rigoureuse, proches de ce que l'on pourrait trouver dans un sujet d'examen réel (Welbl *et al.*, 2017), un exemple de question est présenté dans la table 1. Ces questions sont ouvertes et leur traitement nécessite d'identifier et de comprendre les connaissances scientifiques pertinentes, avant de suivre certains raisonnemment pour y répondre (Yu *et al.*, 2024). Ce type de questions, nous permet d'une part d'évaluer la capacité de récupération des connaissances scientifiques pertinentes, et d'autre part, d'évaluer la capacité de raisonner du modèle. Un autre avantage de SCIQ est le format des questions, à choix multiples, qui permet une évaluation automatique des réponses.

Ici, nous utilisons SCIQ pour comparer deux approches : un modèle génératif classique basé sur un LLM, et un système RAG dans lequel les réponses sont générées à partir d'un ensemble de documents récupérés automatiquement. Nous cherchons à évaluer non seulement la performance quantitative (précision des réponses), mais également la qualité des justifications fournies par le système.

3.2 Modèles et configurations expérimentés

3.2.1 Baseline : un modèle LLM seul

Comme référence de génération de réponse, nous avons choisi d'utiliser les grands modèles de langue (LLM) Llama 3 [8] et Mistral 7b [9] seuls. Ils permettent de générer une réponse aux questions selon une stratégie "sans exemple" (*zero-shot*). Cette stratégie n'exploite pas de source externe d'informations

8. https://www.llama.com/models/llama-3/
9. https://mistral.ai/news/announcing-mistral-7b

mais se base uniquement sur celles acquises lors de l'entraînement du LLM. Le modèle reçoit la question telle quelle et génère directement une réponse. Plus précisément, nous utilisons les modèles *Llama3 8b* et *Mistral 7b*[10]. Un exemple d'invite est donné dans la figure 2 : Invite 1.

3.2.2 Stratégie RAG

L'architecture de notre approche RAG, présentée dans la figure 1, est constituée de :

- **module de récupération d'information (*Retriever*) :** ce module a pour rôle de rechercher et retourner les documents pertinents à partir d'une source externe, ici un corpus documentaire, afin de répondre à une question donnée ;
- **module de génération de la réponse (*Generator*) :** ce module génère une réponse en se basant sur les documents pertinents rassemblés précédemment ;
- **base de connaissance :** cette base a été construite à partir des explications des questions de l'ensemble de données SCIQ. Ces passages explicatifs (tableau 1) jouent le rôle d'un corpus de documents concis et fiables.

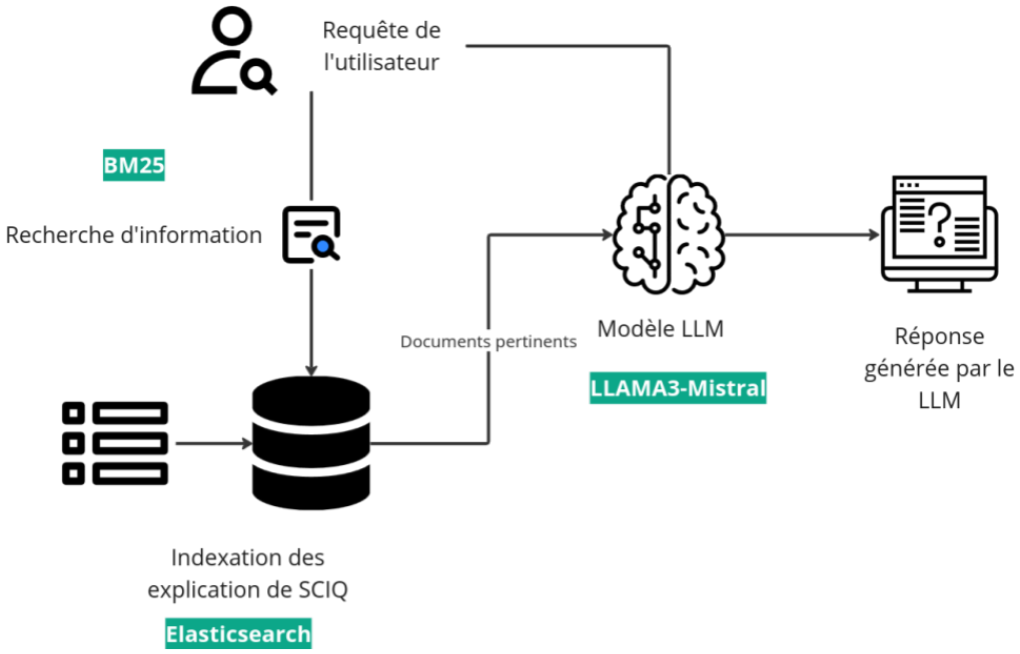

FIGURE 1 – Architecture de l'approche basée sur le RAG : des documents sont indexés puis récupérés par le moteur de recherche ElasticSearch en exploitant une approche BM25, les documents trouvés et la question sont transmises au LLM qui génère alors une réponse.

Nous commençons par construire une base de connaissance. Nous considérons les explications de chaque question de l'ensemble SCIQ comme autant d'unités documentaires à indexer dans le moteur de recherche ElasticSearch[11] avec un identifiant unique. ElasticSearch est une base de données NoSQL orientée documents, optimisée pour la recherche et l'analyse en temps réel, qui nous permet de créer différents types d'index et d'effectuer plusieurs types de recherche.

10. https://ollama.com/library/llama3 et https://ollama.com/library/mistral
11. https://www.elastic.co/fr/elasticsearch

Le système, le module RAG et le modèle de langue de référence (ici soit Llama 3, soit Mistral 7b), reçoivent, de la part de l'utilisateur, uniquement la question du jeu de données SCIQ en entrée. Aucune proposition de réponse (choix multiple) ni explication supplémentaire ne sont fournies au système dans l'invite (*prompt*). Le moteur de recherche documentaire ElasticSearch permet d'identifier des documents supposés pertinents pour une requête en utilisant une approche "sac de mots", avec une fonction de score de type BM25 (Robertson *et al.*, 1995). Les k passages ayant les scores les plus élevés sont retenus. Après avoir testé plusieurs valeurs de k : 1,3,5,10 et 20, nous avons fixé $k = 10$. Ces documents sont ensuite transmis au module de génération de réponse. Les documents retournés par le Retriever sont injectés dans une invite structurée, utilisée pour générer la réponse finale via le LLM. Un exemple d'invite donné au LLM est donné dans la figure 2 : Invite 2. La formulation de l'invite encourage des réponses concises et adaptées au format du jeu de données.

Invite 1 - stragégie de référence LLM seul

Answer the following question with a short and simple response (a few words only).
If you don't know the answer, say 'Not found'.
Question : "question SCIQ"
Answer :

Invite 2 - stratégie RAG

Using the context below, answer the question that follows with a short and simple response (a few words only). If the answer cannot be found in the context, say 'Not found'.
Context : "documents"
Question : "question SCIQ"

FIGURE 2 – Exemple d'invites pour les stratégies LLM seul et RAG.

3.3 Protocole d'évaluation

Evaluation de l'étape de recherche de documents : nous utilisons l'outil standard trec_eval [12] qui permet d'évaluer la pertinence des documents récupérés en fonction d'une liste de référence (*gold standard*, construit ici en association à chaque question sa *bonne* explication telle que donnée dans SCIQ), à l'aide de métriques telles que le *Mean Average Precision (MAP)*, *Mean Reciprocal Rank (MRR)*, la *Precision@k*, le *Rappel@k* et le *Normalized Discounted Cumulative Gain (nDCG)*. Ces mesures permettent d'évaluer non seulement si les documents retournés sont pertinents, mais aussi leur position dans la liste de résultats, ce qui est crucial pour l'efficacité d'un système RAG, où seuls les premiers documents sont utilisés pour générer la réponse finale. Dans notre cas, nous nous concentrons sur le MRR et le Rappel (global et R@k), car notre référence de jugements de pertinence ne contient qu'un seul bon document pertinent par requête.

Evaluation des réponses générées : pour évaluer la qualité des réponses générées par le système, plusieurs métriques complémentaires ont été utilisées. Tout d'abord, une mesure de correspondance exacte (Exact Match) $EM1$ qui permet de vérifier si la réponse du système correspond mot à mot à la

12. https://trec.nist.gov/trec_eval

réponse attendue. Puis $EM2$ pour considérer le cas où la bonne réponse est une sous-chaîne de la réponse générée. Ensuite, la distance de Levenshtein (Navarro, 2001) DL mesure la similarité entre la réponse générée et la réponse de référence en entier et DL_Part (partiel) en tenant compte des sous-chaînes les plus proches, offrant ainsi une évaluation plus tolérante aux variations de formulation. Des métriques classiques telles que la Précision P, le Rappel R et le $F1 - score$ permettent quant à elles d'évaluer les réponses selon une échelle binaire, réponse correcte ou non. Pour aller au-delà de la simple comparaison lexicale, une mesure d'exactitude (*accurracy*) selon un score de similitude sémantique est également appliquée, à l'aide du modèle *paraphrase-MiniLM-L6-v2*[13] $SS1$ et le modèle *all-mpnet-base-v2*[14] $SS2$, afin d'évaluer si le sens général des réponses est proche de celui des références attendues. Enfin, un LLM peut lui-même être utilisé pour estimer la pertinence des réponses générées.

4 Résultats et discussion

4.1 Evaluation de la recherche de documents

Nous avons commencé par évaluer la capacité de notre système RAG à retourner les explications (documents) pertinentes pour les requêtes. Nous avons effectué ce type sur l'ensemble de données *train* de SCIQ (table 2).

Modèle	MRR	Rappel	R@5
RAG-Simple	0,7247	0,8425	0,8175

TABLE 2 – Evaluation de la recherche de documents pertinents

— La valeur élevée du MRR (0,7247) indique que le premier document pertinent est généralement retrouvé parmi les premières positions du classement, ce qui témoigne de la capacité du système à retourner efficacement l'information pertinente dès les premiers résultats.
— Rappel global (0,8425) et R@5 (0,8175) sont tous les deux élevés : ce qui confirme que notre document pertinent est souvent dans le top 5.

4.2 Évaluation de la réponse générée

Nous avons ensuite évalué la qualité des réponses retournées par l'approche RAG en les comparant avec la référence LLM seul (tableau 3). La stratégie `rag-llama` (utilisant Llama3 avec RAG) obtient les meilleurs scores globaux sur la majorité des métriques (EM1, DL, SS1, SS2, Précision, F1) Les approches de base `bl-mist` et `bl-llama` avec les modèles llm Mistral et LLAMA3 utilisés sans RAG ont des performances globalement inférieures. Ces résultats confirment que notre solution basée sur le RAG peut effectivement améliorer la précision de la réponse. Nous notons aussi que le modèle Llama3 est globalement plus performant que Mistral 7b pour SCIQ.

13. https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2
14. https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Stratégie	EM1	EM2	DL	DL_Part	SS1	SS2	P	R	F1	LLM
bl-mist	0,27	0,51	0,29	0,61	0,41	0,36	0,22	0,54	0,31	0,82
bl-llama	0,39	0,46	0,42	0,59	0,51	0,47	0,45	0,47	0,46	0,75
rag-llama	**0,46**	0,56	**0,50**	0,68	**0,59**	**0,55**	**0,49**	0,56	**0,52**	0,79
rag-mist	0,38	**0,74**	0,41	**0,82**	0,54	0,47	0,15	**0,78**	0,25	**0,94**

TABLE 3 – Evaluation des réponses générées, avec RAG (rag) ou sans RAG (bl), avec Llama 3 (llama) ou Mistral 7b (mist).

4.3 Traçabilité des réponses : gestion de l'absence d'information

Nous nous plaçons maintenant dans un cadre de test où les questions posées ne bénéficient pas d'explication réponse explicite dans la base de connaissance (il n'y a pas de *document support*). Nous avons pris les questions de l'ensemble *test* du jeu de données SCIQ car les supports de ces derniers n'ont pas été indexés dans notre base de connaissance (cette dernière contient les supports de l'ensemble *train* seulement). Nous avons interrogé le modèle baseline **Llama3** avec ces question ainsi que notre système rag (représenté par bl-llama et rag-llama dans le tableau 4.3).

Modèle	EM1	EM2	DL	DL_Part	SS1	SS2	P	R	F1	LLM
bl-llama	**0,39**	**0,48**	**0,43**	**0,59**	**0,50**	**0,47**	**0,45**	**0,47**	**0,46**	**0,74**
rag-llama	0,33	0,39	0,35	0,50	0,42	0,39	0,26	0,39	0,31	0,58

TABLE 4 – Evaluation de la réponse générée

— Nous remarquons une diminution très nette pour toutes les métriques du système RAG dans ce cas où les supports des questions ne sont pas présents dans la base de connaissance.
— Nous notons aussi que les métriques de notre baseline n'a pas été affecté par ce type test (comparé avec les résultats du tableau 3). La baseline a surpassé le RAG.

L'influence de l'absence de documents dans un système RAG se manifeste par des réponses plus prudentes ou absentes (dans ce cas le RAG nous a retourné plus de réponses de type "not found"). Dans le contexte éducatif, il est préférable de signaler un manque d'information plutôt que de générer une réponse incorrecte, grâce au RAG, l'hallucination du LLM est réduite.

4.4 Clarté des explications

La stratégie RAG permet de préciser le document support utilisé pour générer la réponse. De plus, nous pouvons indiquer au LLM du RAG d'expliquer la réponse et pourquoi avoir choisie celle-ci. Dans ce cas 4.4, le support utilisé par notre RAG est le même que celui qui devrait être utilisé, tel que mentionné dans l'ensemble de données. Nous avons retrouvé d'autres cas, où la réponse est correcte et le support utilisé par le module de génération est bien pertinent pour répondre à la question mais n'est pas celui attendu, comme dans le jeu de données.

Question	Réponse	support	Explication
Transform faults are the site of massive what ?	earthquakes	ID_support :94 Transform faults are the site of massive earthquakes.	In the provided context, it is stated that “Transform faults are the site of massive earthquakes” (Paragraph 94).
In a glass of sweet tea the sugar is known as the solute and the water is known as what ?	solvent	ID_support :123 If you’re like Tanya in this picture, you prefer your iced tea sweetened with sugar. Sweetened iced tea is a solution in which solid sugar (the solute) is dissolved in cold liquid tea, which is mostly water (the solvent). When you add sugar to tea, particles of water pull apart particles of sugar. The particles of sugar spread throughout the tea, making all of it taste sweet.	In the context, the definition of iced tea (paragraph 123) states that solid sugar is dissolved in cold liquid tea. The soluble substance is the solute and the liquid it is dissolved in is the solvent.

TABLE 5 – Exemples de réponses générées et d’explications associées.

5 Conclusion et perspectives

Plusieurs pistes d’amélioration restent ouvertes. Le jeu de données utilisé est bien connu et largement exploité dans la littérature, ce qui signifie que certains LLMs ont probablement été entraînés sur ce contenu. Cela peut expliquer les performances élevées observées pour les modèles de référence, LLM seuls. La recherche de documents pertinents reste peu performante en termes de classement des documents retrouvés, ce qui suggère la nécessité d’améliorer ce composant, notamment via des techniques de réordonnancement.

Cependant, l’intégration de l’approche RAG dans notre solution a permis d’améliorer la qualité des réponses générées, avec des gains notables notamment avec la stratégie `rag-llama`. Ces résultats confirment la pertinence du RAG dans un système de questions réponses, en particulier dans le contexte éducatif. En effet, dans ce cadre, ces améliorations sont significatives : une meilleure précision et un rappel élevé permettent de fournir aux apprenants des réponses plus fiables et complètes, tout en réduisant les risques liés aux hallucinations des modèles. Ainsi, les systèmes RAG peuvent contribuer à la mise en place d’outils d’assistance à l’apprentissage plus efficaces et mieux adaptés aux besoins spécifiques des élèves.

Références

ABRAHAM S., EWARDS V. & TERENCE S. (2024). Interactive Video Virtual Assistant Framework with Retrieval Augmented Generation for E-Learning. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, p. 1192–1199. DOI : 10.1109/ICAAIC60222.2024.10575255.

AL GHADBAN Y., LU H. Y., ADAVI U., SHARMA A., GARA S., DAS N., KUMAR B., JOHN R., DEVARSETTY P. & HIRST J. E. (2023). Transforming Healthcare Education : Harnessing Large Language Models for Frontline Health Worker Capacity Building using Retrieval-Augmented Generation. DOI : 10.1101/2023.12.15.23300009.

BARUA P. D., VICNESH J., GURURAJAN R., OH S. L., PALMER E., AZIZAN M. M., KADRI N. A. & ACHARYA U. R. (2022). Artificial Intelligence Enabled Personalised Assistive Tools to Enhance Education of Children with Neurodevelopmental Disorders—A Review. *International Journal of Environmental Research and Public Health*, **19**(3), 1192. Number : 3 Publisher : Multidisciplinary Digital Publishing Institute, DOI : 10.3390/ijerph19031192.

BOSCH N. & D'MELLO S. K. (2021). Automatic Detection of Mind Wandering from Video in the Lab and in the Classroom. *IEEE Transactions on Affective Computing*, **12**(4), 974–988. DOI : 10.1109/TAFFC.2019.2908837.

CUI J., NING M., LI Z., CHEN B., YAN Y., LI H., LING B., TIAN Y. & YUAN L. (2024). Chatlaw : A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model. arXiv :2306.16092 [cs], DOI : 10.48550/arXiv.2306.16092.

DEAN M., BOND R. R., MCTEAR M. F. & MULVENNA M. D. (2023). ChatPapers : An AI Chatbot for Interacting with Academic Research. In *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, p. 1–7, Letterkenny, Ireland : IEEE. DOI : 10.1109/AICS60730.2023.10470521.

DEL BONIFRO F., GABBRIELLI M., LISANTI G. & ZINGARO S. P. (2020). Student Dropout Prediction. In I. I. BITTENCOURT, M. CUKUROVA, K. MULDNER, R. LUCKIN & E. MILLÁN, Éds., *Artificial Intelligence in Education*, volume 12163, p. 129–140. Cham : Springer International Publishing. Series Title : Lecture Notes in Computer Science, DOI : 10.1007/978-3-030-52237-7_11.

DEVECI TOPAL A., DILEK EREN C. & KOLBURAN GEÇER A. (2021). Chatbot application in a 5th grade science course. *Education and Information Technologies*, **26**(5), 6241–6265. DOI : 10.1007/s10639-021-10627-8.

DOLENC K., ABERŠEK B. & KORDIGEL ABERŠEK M. (2015). ONLINE FUNCTIONAL LITERACY, INTELLIGENT TUTORING SYSTEMS AND SCIENCE EDUCATION. *Journal of Baltic Science Education*, **14**(2), 162–171. DOI : 10.33225/jbse/15.14.162.

FOROOTANI A., ALIABADI D. E. & THRAEN D. (2025). Bio-Eng-LMM AI Assist chatbot : A Comprehensive Tool for Research and Education. arXiv :2409.07110 [eess], DOI : 10.48550/arXiv.2409.07110.

HABIB M. A., AMIN S., OQBA M., JAIPAL S., KHAN M. J. & SAMAD A. (2024). TaxTajweez : A Large Language Model-based Chatbot for Income Tax Information In Pakistan Using Retrieval Augmented Generation (RAG). *The International FLAIRS Conference Proceedings*, **37**. DOI : 10.32473/flairs.37.1.135648.

HOLMES W. & TUOMI I. (2022). State of the art and practice in AI in education. *European Journal of Education*, **57**(4), 542–570. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/ejed.12533, DOI : 10.1111/ejed.12533.

JIANG Y., SHAO Y., MA D., SEMNANI S. J. & LAM M. S. (2024). Into the Unknown Unknowns : Engaged Human Learning through Participation in Language Model Agent Conversations. arXiv :2408.15232 [cs], DOI : 10.48550/arXiv.2408.15232.

KITTO K., SARATHY N., GROMOV A., LIU M., MUSIAL K. & BUCKINGHAM SHUM S. (2020). Towards skills-based curriculum analytics : can we automate the recognition of prior learning ? In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, p. 171–180, Frankfurt Germany : ACM. DOI : 10.1145/3375462.3375526.

LANE H. C., CAHILL C., FOUTZ S., AUERBACH D., NOREN D., LUSSENHOP C. & SWARTOUT W. (2013). The Effects of a Pedagogical Agent for Informal Science Education on Learner Behaviors and Self-efficacy. In D. HUTCHISON, T. KANADE, J. KITTLER, J. M. KLEINBERG, F. MATTERN, J. C. MITCHELL, M. NAOR, O. NIERSTRASZ, C. PANDU RANGAN, B. STEFFEN, M. SUDAN, D. TERZOPOULOS, D. TYGAR, M. Y. VARDI, G. WEIKUM, H. C. LANE, K. YACEF, J. MOSTOW & P. PAVLIK, Éds., *Artificial Intelligence in Education*, volume 7926, p. 309–318. Berlin, Heidelberg : Springer Berlin Heidelberg. Series Title : Lecture Notes in Computer Science, DOI : 10.1007/978-3-642-39112-5_32.

LEE H., PALLANT A., PRYPUTNIEWICZ S., LORD T., MULHOLLAND M. & LIU O. L. (2019). Automated text scoring and real-time adjustable feedback : Supporting revision of scientific arguments involving uncertainty. *Science Education*, **103**(3), 590–622. DOI : 10.1002/sce.21504.

LEVONIAN Z., LI C., ZHU W., GADE A., HENKEL O., POSTLE M.-E. & XING W. (2023). Retrieval-augmented Generation to Improve Math Question-Answering : Trade-offs Between Groundedness and Human Preference. arXiv :2310.03184 [cs], DOI : 10.48550/arXiv.2310.03184.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T. *et al.* (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, **33**, 9459–9474.

LI Y., LI Z., ZHANG K., DAN R., JIANG S. & ZHANG Y. (2023). ChatDoctor : A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus*. DOI : 10.7759/cureus.40895.

LIN Y.-H. (2019). A Supportive Information Assistant on Mobile Devices for Non-Technical Students Learning Programming.

LIU Y., CAO J., LIU C., DING K. & JIN L. (2024). Datasets for Large Language Models : A Comprehensive Survey. arXiv :2402.18041 [cs], DOI : 10.48550/arXiv.2402.18041.

LÁLA J., O'DONOGHUE O., SHTEDRITSKI A., COX S., RODRIQUES S. G. & WHITE A. D. (2023). PaperQA : Retrieval-Augmented Generative Agent for Scientific Research. arXiv :2312.07559 [cs], DOI : 10.48550/arXiv.2312.07559.

MAESTRALES S., ZHAI X., TOUITOU I., BAKER Q., SCHNEIDER B. & KRAJCIK J. (2021). Using Machine Learning to Score Multi-Dimensional Assessments of Chemistry and Physics. *Journal of Science Education and Technology*, **30**(2), 239–254. DOI : 10.1007/s10956-020-09895-9.

MAITY S. & DEROY A. (2024). Generative AI and Its Impact on Personalized Intelligent Tutoring Systems. arXiv :2410.10650 [cs], DOI : 10.48550/arXiv.2410.10650.

MARCINKOWSKI F., KIESLICH K., STARKE C. & LÜNICH M. (2020). Implications of AI (un-)fairness in higher education admissions : the effects of perceived AI (un-)fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, p. 122–130, Barcelona Spain : ACM. DOI : 10.1145/3351095.3372867.

MAYER R. E. (2019). Computer Games in Education. *Annual Review of Psychology*, **70**(Volume 70, 2019), 531–549. Publisher : Annual Reviews, DOI : 10.1146/annurev-psych-010418-102744.

MCLAREN B. M., ADAMS D. M., MAYER R. E. & FORLIZZI J. (2017). A Computer-Based Game that Promotes Mathematics Learning More than a Conventional Approach :. *International Journal of Game-Based Learning*, **7**(1), 36–56. DOI : 10.4018/IJGBL.2017010103.

NAVARRO G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, **33**(1), 31–88.

NIGAM A., PASRICHA R., SINGH T. & CHURI P. (2021). A Systematic Review on AI-based Proctoring Systems : Past, Present and Future. *Education and Information Technologies*, **26**(5), 6421–6445. DOI : 10.1007/s10639-021-10597-x.

PARONG J., MAYER R. E., FIORELLA L., MACNAMARA A., HOMER B. D. & PLASS J. L. (2017). Learning executive function skills by playing focused video games. *Contemporary Educational Psychology*, **51**, 141–151. DOI : 10.1016/j.cedpsych.2017.07.002.

PASCHOAL L. N., DE OLIVEIRA M. M. & CHICON P. M. M. (2018). A Chatterbot Sensitive to Student's Context to Help on Software Engineering Education. In *2018 XLIV Latin American Computer Conference (CLEI)*, p. 839–848, São Paulo, Brazil : IEEE. DOI : 10.1109/CLEI.2018.00105.

ROBERTSON S. E., WALKER S., JONES S., HANCOCK-BEAULIEU M. M., GATFORD M. *et al.* (1995). Okapi at trec-3. *Nist Special Publication Sp*, **109**, 109.

SREELAKSHMI A., ABHINAYA S., NAIR A. & JAYA NIRMALA S. (2019). A Question Answering and Quiz Generation Chatbot for Education. In *2019 Grace Hopper Celebration India (GHCI)*, p. 1–6, Bangalore, India : IEEE. DOI : 10.1109/GHCI47972.2019.9071832.

SUNG S. H., LI C., CHEN G., HUANG X., XIE C., MASSICOTTE J. & SHEN J. (2021). How Does Augmented Observation Facilitate Multimodal Representational Thinking ? Applying Deep Learning to Decode Complex Student Construct. *Journal of Science Education and Technology*, **30**(2), 210–226. DOI : 10.1007/s10956-020-09856-2.

SWACHA J. & GRACEL M. (2025). Retrieval-Augmented Generation (RAG) Chatbots for Education : A Survey of Applications. *Applied Sciences*, **15**(8), 4234. Number : 8 Publisher : Multidisciplinary Digital Publishing Institute, DOI : 10.3390/app15084234.

WARD W., COLE R., BOLAÑOS D., BUCHENROTH-MARTIN C., SVIRSKY E. & WESTON T. (2013). My science tutor : A conversational multimedia virtual tutor. *Journal of Educational Psychology*, **105**(4), 1115–1125. DOI : 10.1037/a0031589.

WELBL J., LIU N. F. & GARDNER M. (2017). Crowdsourcing Multiple Choice Science Questions. arXiv :1707.06209 [cs], DOI : 10.48550/arXiv.1707.06209.

WIRATUNGA N., ABEYRATNE R., JAYAWARDENA L., MARTIN K., MASSIE S., NKISI-ORJI I., WEERASINGHE R., LIRET A. & FLEISCH B. (2024). CBR-RAG : Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering. arXiv :2404.04302 [cs], DOI : 10.48550/arXiv.2404.04302.

YU H.-C., SHIH Y.-A., LAW K.-M., HSIEH K.-Y., CHENG Y.-C., HO H.-C., LIN Z.-A., HSU W.-C. & FAN Y.-C. (2024). Enhancing Distractor Generation for Multiple-Choice Questions with Retrieval Augmented Pretraining and Knowledge Graph Integration. arXiv :2406.13578 [cs], DOI : 10.48550/arXiv.2406.13578.