Cadre d'évaluation pour les systèmes de génération augmentée (RAG) : combinaison des performances de recherche d'informations et de LLM

Mohamed-Amine El-Yagouby[1, 2] Philippe Mulhem[1]
Jean-Pierre Chevallet[1] Eric Gaussier[1]
(1) LIG UGA, Bâtiment IMAG, 700 Av. Centrale, 38401 Saint-Martin-d'Hères
(2) LORIA, Campus Scientifique, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy
mohamed-amine.el-yagouby@loria.fr,
philippe.mulhem@univ-grenoble-alpes.fr
jean-pierre.chevallet@univ-grenoble-alpes.fr,
eric.gaussier@univ-grenoble-alpes.fr

RÉSUMÉ

Cet article introduit un nouveau cadre d'évaluation pour les systèmes RAG, en comblant les lacunes des approches précédentes. La première phase consiste à concevoir un ensemble de données avec des parties pertinentes extraites pour chaque exemple, représentant les informations nécessaires pour répondre à une question donnée, et à proposer une métrique d'évaluation pour les systèmes IR basée sur la présence de ces parties dans le contenu récupéré. La deuxième phase explore la relation entre le système de RI et les évaluations RAG globales et utilise cette relation pour prédire les performances globales du RAG à partir des performances du SRI. Cette approche élimine le besoin de réponses coûteuses générées par LLM et d'évaluations ultérieures, réduisant ainsi les coûts et fournissant un cadre d'évaluation plus complet et plus robuste pour les systèmes RAG.

ABSTRACT

Evaluation framework for Retrieval-Augmented Generation (RAG) systems : combining information retrieval and LLM performance.

This work introduces a new evaluation framework for RAG systems, which addresses the shortcomings of previous approaches. The first phase involves designing a dataset with relevant parts extracted for each example, representing the necessary information to answer a given question, and proposing an evaluation metric for IR systems based on the presence of these parts in the retrieved content. The second phase explores the relationship between IR and overall RAG evaluations and uses this relationship to predict the overall RAG performance from its retrieval component performance. This approach eliminates the need for expensive LLM-generated responses and subsequent evaluations, reducing costs and providing a more comprehensive and robust evaluation framework for RAG systems.

MOTS-CLÉS : RAG, cadre, évaluation.

KEYWORDS: RAG, framework, evaluation.

ARTICLE : **Accepté à** CORIA 2025.

1 Introduction

Les approches RAG (Retrieval Augmented Generation) combinent les capacités des grands modèles de langue (LLM) en matière de compréhension et de génération de langage avec les systèmes de recherche d'informations (SRI) pour générer des réponses plus précises et contextuellement pertinentes (Gao *et al.*, 2024). Le processus typique d'un RAG consiste à récupérer des documents pertinents sur la base d'une requête utilisateur et d'un SRI et à utiliser ces documents pour inciter les LLM à générer des réponses de meilleur qualité. Les approches RAG peuvent être utilisés au-delà des applications de questions-réponses (Jiang *et al.*, 2023), et ont démontré leur efficacité dans la réduction des hallucinations (Zhang *et al.*, 2023), l'ancrage des connaissances à partir de sources multiples (Gao *et al.*, 2024) et, surtout, l'accès à des informations dynamiques et actualisées (Gao *et al.*, 2024).

L'évaluation des RAG peut être difficile : contrairement aux modèles de langage traditionnels, les performances d'un système RAG dépendent non seulement de la qualité du texte généré, mais également de la qualité des documents récupérés. Par conséquent, les mesures d'évaluation doivent tenir compte de ces deux aspects. Les mesures d'évaluation classiques d'un SRI sont la MAP, le score F1, le MRR et le NDCG qui évaluent la capacité du système à récupérer les documents pertinents en réponse à une requête. Les mesures traditionnelles de génération de langage naturel (NLG) telles que les scores BLEU et ROUGE peuvent être utilisées pour mesurer la pertinence du texte généré des LLM. L'évaluation humaine joue un rôle crucial dans l'évaluation des LLM, en fournissant des informations sur des aspects tels que la cohérence, la fluidité et la qualité globale que les mesures automatisées pourraient manquer. Pour l'évaluation globale d'un RAG, certaines des mesures précédentes sont utilisées, telles que ROUGE (Salemi & Zamani, 2024), et d'autres techniques basées sur l'emploi d'un LLM comme juge pour calculer les scores qui évaluent la qualité du système RAG sur différents aspects tels que la fidélité des réponses (Saad-Falcon *et al.*, 2024) et la pertinence du contexte (Es *et al.*, 2024).

La question de recherche traitée dans cet article est : **comment évaluer les systèmes de recherche d'informations (SRI) dans le cadres de génération augmentée par LLM (RAG) ?** Cette question est traitée en trois sous-questions :

1. **Comment différents systèmes de recherche d'informations impactent les performances globales d'un système RAG ?**

   L'évaluation des réponses générées qui représentent les performances globales d'un système RAG avec plusieurs systèmes de recherche d'informations peut révéler l'influence de ces systèmes sur la précision et la qualité des réponses générées.

2. **Quelle est la relation entre la qualité du système de recherche d'informations et la qualité globale d'un système RAG ?**

   La compréhension de cette relation peut nous aider à répondre à la question suivante et ajoute de l'explicabilité aux réponses générées.

3. **Pouvons-nous estimer les performances globales d'un système RAG en fonction des performances de son système de recherche d'informations ?**

   L'établissement d'une telle estimation peut nous aider à créer un nouveau cadre d'évaluation pour les systèmes RAG qui soit plus explicable et systématique. Ce cadre évaluerait les deux phases d'un système RAG : la phase de recherche d'informations et la phase de génération. Ce faisant, il fournit une compréhension globale de la manière dont chaque phase contribue à la

performance finale, permettant des améliorations ciblées et des évaluations plus transparentes.

Dans la suite de cet article, nous présentons en partie 2 un état de l'art lié à notre problématique. En partie 3 nous décrivons notre approche. Les expérimentations et les résultats sont présentés en partie 5, avant de conclure.

2 État de l'art

Nous décrivons ici les différentes propositions pour l'évaluation des RAG. RAGAS (Retrieval Augmented Generation Assessment) (Es *et al.*, 2024) fournit une méthode d'évaluation des systèmes RAG sans s'appuyer sur des réponses de référence. Il se concentre sur trois dimensions clés : la fidélité, la pertinence de la réponse et la pertinence du contexte. RAGAS utilise un autre LLM comme juge pour évaluer ces dimensions. La fidélité indique si la réponse tient compte des documents retrouvés. La pertinence de la réponse est mesurée en invitant le LLM évalué à générer des questions potentielles en fonction de la réponse, puis en calculant la similarité entre ces questions et la question d'origine. La pertinence du contexte évalue l'orientation et la nécessité du contexte récupéré (les documents retrouvés par le SRI) pour répondre à la question. Le LLM extrait du contexte les phrases essentielles pour répondre à la question, et le score de pertinence est la proportion de phrases nécessaires dans le contexte. Avec ces mesures, aucune évaluation du SRI n'est donc réalisée directement lors du calcul de le pertinence du contexte.

ARES (Zhang *et al.*, 2023) propose d'automatiser le processus d'évaluation sur les critères de RAGAS. Il aborde le problème en trois étapes : génération de données synthétiques, réglage fin des juges LLM, et enfin inférence basée sur la prédiction. Dans l'étape de génération de données synthétiques, un LLM est invité à générer des questions synthétiques et des réponses correspondantes en fonction de passages fournis. Des LLM légers, tels que DeBERTa-v3-Large, sont ensuite affinés en tant que juges LLM à l'aide de ces données synthétiques pour les trois tâches de classification associées chacune à la mesure d'évaluation : pertinence du contexte, fidélité des réponses, et enfin pertinence des réponses. On se replace ensuite dans le même cadre que des évaluations de RAGAS. Cette approche ne se base donc pas dans un cadre où un SRI est évalué.

D'autres approches, comme CRAG (Meta, 2025), proposée par l'entreprise Meta, utilise cinq domaines de référence : Finance, Sports, Musique, Cinéma et une encyclopédie à domaine ouvert (Wikipedia). Il comporte huit types de questions : questions simples, conditionnelles, ensemble, comparaison, agrégation, multi-sauts, post-traitement et fausses prémisses. CRAG mesure la qualité de prédiction sur une échelle à 3 valeurs : parfait, acceptable et manquant. Cependant, CRAG ne se préoccupe pas spécifiquement de l'évaluation du SRI dans le RAG.

Le benchmark de génération augmentée de récupération (RGB) (Chen *et al.*, 2024) évalue les performances des LLM sur des tâches qui combinent la récupération et la génération, en se concentrant sur quatre capacités principales : la robustesse au bruit, le rejet négatif, l'intégration des informations et la robustesse contrefactuelle. RGB lui non plus n'évalue aucunement la composante SRI.

Notre étude de l'état de l'art, dont un bref rappel a été donné ci-dessus, nous montre qu'il n'y existe pas d'étude explicite sur l'influence de la qualité du SRI sur le RAG. Dans la partie suivante, nous faisons donc une proposition dans ce sens.

3 Proposition

Ce travail vise à évaluer les SRI dans un cadre RAG et à évaluer les performances globales du système RAG, en se concentrant particulièrement sur la qualité de ses réponses finales. De plus, nous explorerons la relation entre ces évaluations et la possibilité d'estimer les performances globales d'un cadre RAG en fonction de ses performances de recherche d'information.

Pour cela, notre cadre d'évaluation utilise un corpus de documents, et est composé de triplets $< Q, A, P >$ formés de :

— une requête Q ;
— la réponse pertinente à la requête Q, notée A ;
— l'ensemble des passages pertinents des documents du corpus pour Q, noté $P = \{p_1, ..., p_n\}$

3.1 Evaluation de la partie SRI

Un LLM est contraint par sa taille limitée de $tokens$ en entrée, notée L, par exemple 4096 pour gpt-3.5-turbo-instruct [1]. Cette limitation a deux impacts dans notre cadre :

— nous choisissons de diviser chaque document en courts passages, afin de garantir que le prompt du RAG soit totalement traité ;
— comme un SRI peut renvoyer des textes (passages) dont la totalité est plus grande que L, nous considérons uniquement une partie du texte total renvoyé par le SRI, en se limitant à aux C_N premiers *token* de la concaténation textuelle des passages retrouvés, notés C.

La métrique d'évaluation proposée pour le SRI est une estimation de la probabilité que chaque passage pertinent p_i soit contenu dans C, par le rapport de la longueur de la plus longue sous-séquence commune LCS de p_i dans C à la longueur de p_i. Pour le score total, nous prenons la moyenne de ces rapports pour toutes les parties :

$$p_i \in P : Score_{ir} = \frac{1}{|P|} . \sum_{i=1}^{|P|} \frac{len(LCS(p_i, C))}{len(p_i)}$$

avec :

— $LCS(X, Y)$ représente la plus longue sous-séquence commune de X dans Y ;
— $len(X)$ désigne la longueur de la chaîne X en nombre de caractères.

Malgré la simplicité de cette approche, elle reste efficace car elle nous permet de prendre en compte les correspondances partielles, fournissant ainsi une évaluation plus nuancée des performances du système de RI qu'une mesure d'inclusion binaire stricte. En considérant les correspondances partielles via le LCS, cette mesure évalue efficacement la capacité du système RI à récupérer des informations pertinentes. De plus, ce score est égal à 1 si toutes les parties nécessaires dans l'ensemble P sont trouvées dans les morceaux récupérés et vers 0 si aucun passage pertinent n'est retrouvé.

3.2 Evaluation de la partie LLM

L'évaluation de la partie LLM est réalisée en se basant sur des prompts posés à GPT-4-o afin d'avoir une échelle à 5 valeurs. Cette approche est inspirée de RAGAS (Es *et al.*, 2024) et ARES (Zhang

1. https://platform.openai.com/docs/models

et al., 2023) vu en partie 2, avec pour objectif d'éviter les évaluations humaines. Ce prompt est le suivant :

```
Task: Assess the given candidate's answers based on the
true answer, references, and using the scoring criteria.
Return only the scores separated by commas.
Question: question
True Answer: true answer
References: references
Candidate's Answers: candidates answers
Scoring Criteria:
If the answer says that there is not enough information
in documents to answer the question, the score is 1.
If the answer is partially correct but in details you
found statements that are incorrect according to the
References, the score is 2.
If the answer is partially correct but it doesn't
completely answer the question due to lack of
information in the documents, the score is 3.
If the answer is fully incorrect, the score is 4.
If the answer is fully correct, the score is 5.
```

4 Expérimentations

4.1 Corpus

Le corpus utilisé lors de nos expérimentations est le *distractor setting* de HotpotQA [2]. Dans ce corpus, un système de questions-réponses lit 10 paragraphes pour fournir une réponse à une question. Il contient un total de 7 404 exemples, chaque exemple est composé de la question, de la réponse et des faits justificatifs représentés par leur document de référence et leur texte. Le corpus de texte contient 100 000 documents d'un dump de Wikipedia de 2017.

4.2 Découpage des passages

La méthode de découpage utilisée est la "Semantic Chunking" [3]. Au lieu de découper le texte avec une taille de bloc fixe, l'algorithme sélectionne de manière adaptative les points d'arrêt entre les phrases en utilisant un seuil de similarité. Un passage contient alors des phrases qui sont normalement sémantiquement liées les unes aux autres.

2. https://hotpotqa.github.io/
3. https://docs.llamaindex.ai/en/stable/examples/node_parsers/semantic_chunking/

4.3 Le SRI et le LLM testés

Dans notre expérience, nous avons utilisé Llama3-8b, un LLM open source de Meta, qui contient 8 milliards de paramètres.

Pour la partie recherche d'informations, nous avons utilisé 5 systèmes différents avec leurs stratégies d'indexation :

1. BM25 : un modèle (non dense) qui estime la pertinence des documents par rapport à une requête en prenant en compte la fréquence des termes, la fréquence inverse des documents et la normalisation de la longueur des documents. Nous avons utilisé PyTerrier4.
2. SIM (Embeddings Similarity Retrieval) : une méthode de recherche dense qui vise à capturer les relations sémantiques entre les mots et les phrases.
3. MultiQuery Retriever (MLQ) : ce sytème utilise un LLM pour générer plusieurs requêtes à partir de points de vues différents pour une requêteutilisateur donnée. Pour chaque requête, il récupère alors un ensemble de documents pertinents et prend l'union de toutes les requêtes pour obtenir un ensemble plus large de documents potentiellement pertinents.
4. MMR (Pertinence marginale maximale) : une méthode qui vise à équilibrer la pertinence et la nouveauté, en garantissant que le contenu des documents récupérés est non seulement pertinent mais également diversifié, réduisant ainsi la redondance.
5. Réo (Réorganisation du contexte long) : une méthode qui réorganise les documents récupérés à l'aide d'une autre recherche qui est dans notre exemple l'Embeddings Similarity Retrieval (SIM) ceci pour éviter une dégradation des performances.

5 Résultats

5.1 Evaluation des SRI

Nous avons exécuté 7404 requêtes afin de récupérer leurs passages pertinents pour chaque SRI. Les fragments récupérés ont été concaténés en un seul texte, et nous avons évalué ce texte en calculant le score défini en partie pour les N premiers token de ce texte, avec N dans 100, 200, . . . , 1000.

N	BM25	sim	mmr	mlq	reo
100	0.40	0.34	0.34	0.29	0.16
200	0.46	0.39	0.38	0.33	0.19
300	0.49	0.41	0.39	0.35	0.20
400	0.52	0.44	0.41	0.37	0.21
500	0.55	0.46	0.42	0.39	0.22
600	0.57	0.47	0.43	0.40	0.23
700	0.58	0.48	0.44	0.41	0.24
800	0.60	0.50	0.44	0.42	0.25
900	0.61	0.51	0.45	0.43	0.26
1000	0.62	0.52	0.46	0.43	0.26

TABLE 1 – Moyennes des scores RI d'évalution des SRI avec les premier N tokens.

Pour tous les SRI, les scores RI augmentent quand le nombre de token N augmente. Cela indique qu'un passage plus long fournit des informations plus pertinentes, améliorant ainsi la qualité de la recherche.

On remarque que BM25 surpasse systématiquement les autres SRI sur toutes les longueurs considérées, ce qui indique qu'il s'agit de la meilleure méthode parmi celles évaluées.

L'amélioration des scores diminue au fur et à mesure que la longueur du token augmente. Par exemple, le score de BM25 augmente davantage entre 100 à 500 tokens qu'entre de 500 et 1000 tokens. Cela suggère des résultats décroissants pour les passages plus longs, ce qui signifie que l'ajout de tokens supplémentaires n'apporte de moins en moins d'avantage en termes de score RI. Cela est dû au fait que le texte peut devenir redondant ou que les informations ajoutées peuvent ne pas être aussi pertinentes que les informations initiales. Les résultats décroissants observés suggèrent que des recherches ultérieures pourraient se concentrer sur l'amélioration de la qualité de la recherche sans simplement augmenter la longueur du passage, par exemple en améliorant la pertinence des passages récupérés ou en optimisant les processus de concaténation et de résumé. Ce qui est important compte tenu des limitations de taille d'entrée des LLM et du problème de «Lost in the Middle» (Liu *et al.*, 2024), où des informations importantes au milieu de longs textes peuvent recevoir moins d'attention. Il est donc crucial d'identifier une longueur de token N optimale qui équilibre la qualité de la recherche et les contraintes pratiques.

5.2 Evaluation des LLM

Suite aux résultats précédents, nous avons choisi de fixer $N = 1000$ pour les expérimentations suivantes. Les résultats de l'évaluation utilisant GPT4-o pour l'évaluation du composant générateur (LLM) avec différents systèmes RI dans le cadre RAG du système RAG global en utilisant chacun des 5 systèmes de récupération avec N=1000 sont présentés dans la table 2. Cette table contient aussi le résultat en utilisant uniquement le LLM (*LLM-only*) pour comparaison. Ce tableau présente le pourcentage d'exemples avec leurs scores correspondants (de 1 à 5) est présenté dans cette table.

SRI	1	2	3	4	5
LLM-only	0.01	0.13	0.04	0.58	0.24
BM25	0.08	0.05	0.13	0.31	0.42
sim	0.09	0.05	0.14	0.40	0.32
mmr	0.10	0.07	0.12	0.42	0.29
mlq	0.18	0.06	0.12	0.41	0.24
reo	0.29	0.05	0.07	0.41	0.18

TABLE 2 – Pourcentage des scores LLM de l'évalution à 5 valeurs, avec $N = 1000$ tokens.

Les résultats montrent que l'utilisation du LLM sans aucun système de RI entraîne des problèmes importants d'hallucinations, comme en témoignent les pourcentages élevés dans les scores 2 et 4. Cela souligne le rôle essentiel des systèmes de RI dans l'amélioration de l'exactitude des informations dans le contenu généré.

Les performances des systèmes RAG utilisant chacun des cinq systèmes RI sont fortement corrélées aux performances individuelles de ces systèmes RI, comme évalué précédemment dans la section 4.1. Leur ordre sur le pourcentage élevé de réponses entièrement correctes (score 5) et le faible

pourcentage d'hallucinations et de réponses incorrectes (scores 2 et 4) correspond à l'ordre de leurs évaluations précédentes. Le SRI à base de BM25 apparaît comme le plus efficace parmi les autres SRIs, fournissant des réponses précises et fiables avec des hallucinations relativement minimes.

Le LLM sans SRI répond presque toujours à la question, correctement ou incorrectement, et indique rarement un manque d'informations dans la réponse (seulement 1% + 4% pour les scores 1 et 3). Cela se traduit par des taux élevés d'hallucinations (58% + 13% pour les scores 4 et 2). Même avec des systèmes RI moins performants comme reo, le LLM dans cette configuration RAG peut indiquer des informations insuffisantes (29% + 7% pour les scores 1 et 3), montrant qu'il n'est pas capable de trouver des informations pertinentes pour répondre à certaines requêtes, et évite davantage les hallucinations que le LLM seul.

5.3 Estimation des performances globales du système RAG à partir des performances de son SRI

Comme indiqué dans la section précédente, les performances du système RAG semblent corrélées aux performances du SRI utilisé. Nous détaillons notre analyse par la figure 1. Cette figure présente 3 courbes, toutes relatives à l'utilisation de BM25 comme SRI : chaque courbe présente les requêtes triées par ordre de score de RI croissant avec le score de LLM de 5, donc une bonne réponse (courbe en haut), avec un score LLM de 2, 3, ou 4, donc un réponse partielle (courbe en bas à gauche), et un score de 1 avec une réponse mauvaise (courbe en bas à droite). Nous trouvons donc que les réponses avec des scores RI élevés ont tendance à avoir la plus grande précision (quand le score LLM = 5). À l'inverse, les réponses avec des scores RI faibles reflètent souvent un manque d'informations (score LLM = 1). Et les réponses avec des scores RI intermédiaires contiennent fréquemment des hallucinations ou des informations incomplètes (scores LLM = 2, 3, 4). Cela suggère que lorsque le système RI ne récupère que des informations partielles, le système RAG est plus susceptible de produire un contenu halluciné.

A partir de ces informations, nous présentons une estimation des performances globales d'un RAG à partir de celle du SRI qu'il utilise. Notre méthodologie d'estimation va reposer sur l'estimation de deux seuils, h et k, où :

1. Un score RI inférieur à h indique une forte probabilité que le contenu produise une réponse reflétant un manque d'information (score LLM de 1).
2. Un score RI supérieur à k suggère une forte probabilité de générer une réponse entièrement correcte (score LLM de 5).
3. Un score RI compris entre h et k suggère une forte probabilité de générer des réponses partiellement ou totalement incorrectes avec hallucination (score LLM de 2,3,4).

Pour cela, nous avons appliqué la méthode d'estimation de vraisemblance maximale (MLE). Le processus d'estimation consiste à maximiser la fonction de vraisemblance pour chaque seuil h et k. La formulation pour estimer k est la suivante :

$$\mathcal{L}(k) = -\sum log(p_{ir}.p_{lm} + (1 - p_{ir}).(1 - p_{lm}))$$

avec $p_{LLM} = \begin{cases} 1 & \text{si } score_{LLM} = 5 \\ 0 & \text{sinon} \end{cases}$

et $p_{ir} = \begin{cases} 1 & \text{si } score_{ir} > k \\ 0 & \text{si } score_{ir} \leq k \end{cases}$

La fonction de vraisemblance $\mathcal{L}(k)$ représente la probabilité d'observer un score RI donné supérieur à k et un score LLM de 5, en utilisant les données de l'évaluation automatique des réponses du LLM.

La formulation pour estimer h est similaire à celle pour k :

$$\mathcal{L}(h) = -\sum log(p_{ir}.p_{lm} + (1 - p_{ir}).(1 - p_{lm}))$$

avec p_{LLM} décrit plus haut et $p_{ir} = \begin{cases} 1 & \text{si } score_{ir} < h \\ 0 & \text{si } score_{ir} \geq h \end{cases}$

Pour estimer numériquement ces deux paramètres, avec 5 SRI différents, nous utilisons 37020 (=7404 × 5) couples <score RI, score LLM>. La maximisation de vraisemblance sur les deux paramètres obtient $h = 0.105$ et $k = 0.670$.

Nous présentons en figure 1 les scores RI croissants pour les 7404 requêtes, et les seuils h et k pour le modèle BM25. Il est donc possible par nos travaux, dans une certaine mesure, de prévoir quand le LLM risque d'halluciner.

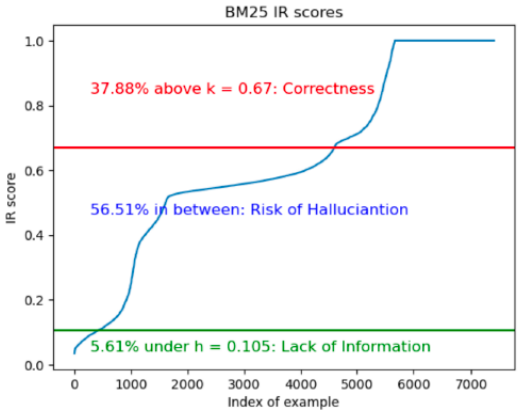

FIGURE 1 – Distribution des scores croissants du SRI avec BM25, avec les seuils k et h

6 Conclusion

Nous avons défini dans cet article un cadre d'évaluation pour un RAG. Notre cadre d'évaluation proposé décompose l'évaluation des performances en deux phases distinctes : la phase de recherche d'informations et la phase de génération. Ce cadre propose une évaluation plus détaillée que les approches existantes de type RAG. Il intègre en particulier des contraintes pratiques imposées par les limitations de taille d'entrée des grands modèles de langage (LLM). Cette approche en deux

phases permet une analyse précise de l'impact des systèmes RI sur la qualité et la précision des réponses générées. Les expériences que nous avons menées révèlent que les différents systèmes RI varient en efficacité, le modèle BM25 surpassant systématiquement les autres en termes de qualité de recherche sur différentes longueurs de textes. Cette étude conclut en fournissant une méthode d'estimation permettant de prédire les performances globales du RAG en fonction des performances de son composant RI.

Dans le futur, nous allons travailler sur l'extension de notre cadre pour intégrer d'autres configurations de RAG, par exemple celles basées sur des approches utilisant des graphes de connaissances.

Ce travail a été mené au Laboratoire d'Informatique de Grenoble (LIG) de l'Université Grenoble Alpes (UGA) et a été financé par le projet ANR GUIDANCE (https://guidance.anr.isir.upmc.fr).

Références

CHEN J., LIN H., HAN X. & SUN L. (2024). Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38 de *AAAI'24/IAAI'24/EAAI'24*, p. 17754–17762 : AAAI Press. DOI : 10.1609/aaai.v38i16.29728.

ES S., JAMES J., ESPINOSA ANKE L. & SCHOCKAERT S. (2024). RAGAs : Automated evaluation of retrieval augmented generation. In N. ALETRAS & O. DE CLERCQ, Éds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, p. 150–158, St. Julians, Malta : Association for Computational Linguistics.

GAO Y., XIONG Y., GAO X., JIA K., PAN J., BI Y., DAI Y., SUN J., WANG M. & WANG H. (2024). Retrieval-Augmented Generation for Large Language Models : A Survey. arXiv :2312.10997 [cs], DOI : 10.48550/arXiv.2312.10997.

JIANG Z., XU F., GAO L., SUN Z., LIU Q., DWIVEDI-YU J., YANG Y., CALLAN J. & NEUBIG G. (2023). Active retrieval augmented generation. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 7969–7992, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.495.

LIU N. F., LIN K., HEWITT J., PARANJAPE A., BEVILACQUA M., PETRONI F. & LIANG P. (2024). Lost in the middle : How language models use long contexts. *Transactions of the Association for Computational Linguistics*, **12**, 157–173.

META (2025). Crag : Comprehensive rag benchmark.

SAAD-FALCON J., KHATTAB O., POTTS C. & ZAHARIA M. (2024). ARES : An automated evaluation framework for retrieval-augmented generation systems. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 338–354, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.20.

SALEMI A. & ZAMANI H. (2024). Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, p. 2395–2400, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3626772.3657957.

ZHANG Y., LI Y., CUI L., CAI D., LIU L., FU T., HUANG X., ZHAO E., ZHANG Y., CHEN Y., WANG L., LUU A. T., BI W., SHI F. & SHI S. (2023). Siren's song in the ai ocean : A survey on hallucination in large language models.