Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Évaluation de métriques de RAG dans un contexte applicatif : une expérience, ses conclusions et ses limites

Quentin Brabant
Orange Research, Lannion, France
quentin.brabant@orange.com

RÉSUMÉ

Cet article rapporte une expérience visant à évaluer la pertinence de différentes métriques de RAG. L'expérience s'appuie sur un jeu de données de type question-réponse, créé par des annotateurices à partir de données métier. Les réponses produites et les passages retrouvés par un système de RAG sont notés à l'aide de métriques d'évaluation du RAG issues de quatre bibliothèques (Ragas, DeepEval, RAGEval, Opik). Ces métriques sont comparées à des notes données par deux évaluateurices, ainsi qu'à des métriques standards telles que le rappel. Une analyse des corrélations est effectuée. Enfin, nous mettons en évidence certaines limites de notre méthodologie, que nous comparons à celles employées dans la littérature, afin de suggérer quelques pistes de recherche.

ABSTRACT

Evaluating RAG Metrics in Applied Contexts : an Empirical Study, its Results and its Limitations

This paper reports on an empirical study for evaluating the relevance of several RAG metrics. The experiment is based on a question-answering dataset created by human annotators from business data. The responses generated and the passages retrieved by a RAG system are scored using evaluation metrics from four libraries (Ragas, DeepEval, RAGEval, Opik). These metrics are compared to scores given by two evaluators, as well as to standard metrics such as recall. An analysis of correlations is conducted. Finally, we highlight certain limitations of our methodology, compare them to those used in the literature, and suggest some avenues for future research.

MOTS-CLÉS : RAG, évaluation, métrique, llm-as-a-judge.

KEYWORDS: RAG, evaluation, metric, llm-as-a-judge.

1 Introduction

Evaluer et comparer des systèmes de RAG (*Retrieval Augmented Generation*) est encore aujourd'hui une tâche difficile : même lorsque l'on dispose d'un ensemble suffisamment grand de questions de test associées à leurs réponses de référence, évaluer automatiquement les réponses d'un système par rapport à ces références n'a rien de trivial. Une approche populaire consiste à utiliser des métriques dites LLM-as-a-judge pour réaliser cette évaluation. Bien que ces métriques semblent généralement plus pertinentes que des métriques classiques telles que BLEU, il est difficile de savoir à l'avance quelle sera la pertinence d'une métrique spécifique sur le jeu de données considéré, d'autant que les critères d'évaluation peuvent varier (pertinence, factualité, complétude des réponses, etc). Il est donc utile, lors de l'évaluation d'un système de RAG en développement, de réaliser une évaluation des

Actes de l'atelier sur l'évaluation des modèles génératifs (LLM) et challenges (EvalLLM 2026), pages 87–98.

métriques disponibles afin de vérifier qu'elles donnent des approximations acceptables du critère considéré, et qu'elles permettront ainsi de comparer de manière fiable les différentes itérations du système de RAG développé. Généralement, les métriques sont évaluées en mesurant leur corrélation à des notes données par des humain·e·s.

Cet article rapporte une expérience de ce type. Cette expérience s'appuie sur un jeu de données de type question-réponse, créé par des annotateurices à partir de données métier. Les réponses produites et les documents retrouvés par un système de RAG sont notées à l'aide de métriques de RAG issues de quatre bibliothèques : Ragas [1](Es *et al.*, 2024), DeepEval [2], RAGChecker [3](Ru *et al.*, 2024), et Opik [4]. Ces métriques sont comparées à des évaluations de références : évaluations humaines pour l'évalsuation des réponses générées, et rappel pour l'évaluation du retrieval.

Notez que notre objectif n'est pas de départager les différentes métriques évaluées, car les résultats de l'étude sont fortement dépendants des choix effectués lors de sa conception. Les expériences rapportées visent plutôt à appliquer une méthodologie afin d'en éprouver les avantages et les limites. Enfin, le jeu de questions-réponses utilisé ne peut malheureusement pas être rendu public ; nous partageons en revanche les scores bruts et le code correspondant aux analyses statistiques effectuées [5].

L'article est organisé de la manière suivante. La section 2 de l'article décrit le contexte d'application et les données utilisées, y compris les processus d'annotation et d'évaluation manuels. La section 3 décrit la méthodologie appliquée afin de comparer les évaluations de référence aux évaluations produites par les métriques issues des bibliothèques testées. La section 4 rapporte et analyse les résultats de cette expérience. Certaines limites de notre méthodologie sont mises en évidence dans la section 5. Dans la section 6, notre méthodologie est comparée à celles employées dans la littérature. Nous proposons finalement quelques perspectives de recherche visant à faciliter l'application de méthodologies fiables pour l'évaluation des métriques de RAG, section 7.

2 Contexte et données

Notre entreprise développe une solution de RAG destinée à répondre à des questions en français portant sur un domaine métier. Le système développé traite chaque question donnée via deux modules clés : d'abord, un *retriever*, dont le rôle est de retrouver des passages pertinents au sein d'une base de documents métier (en français également), qui combine une approche dense avec BM25 ; ensuite un *générateur*, qui injecte la question de l'utilisateur et les 5 premiers passages issus du retriever dans un prompt destiné à GPT 3.5, afin que ce dernier génère la réponse à la question de l'utilisateur. Le retriever est un système hybride s'appuyant à la fois sur une approche dense et sur BM25.

Dans le but d'évaluer la performance de ce système, un dataset a été créé à partir du corpus de documents métiers utilisés par le système de RAG. Ce dataset est un jeu de questions-réponses accompagnés de passages de référence.

1. https://docs.ragas.io/
2. https://docs.confident-ai.com/
3. https://github.com/amazon-science/RAGChecker
4. https://www.comet.com/site/products/opik/
5. https://github.com/Orange-OpenSource/evalllm2026-metric-correlation-analysis

2.1 Le jeu de questions-réponses

Ce jeu de données est composé de 96 questions. Chaque question est associée à une réponse de référence, ainsi qu'à un ou plusieurs passages issus des documents métiers ; ces passages contiennent l'information nécessaire pour répondre à la question à laquelle ils sont associés.

Le jeu de questions-réponses est créé en s'appuyant sur la base de documents métiers existante portant sur le domaine de la télécommunication, et contenant en particulier des informations sur différentes offres ainsi que la relation client. Celle-ci contient 479 documents, de longueur 11 à 14 025 mots (1112 en moyenne). Les questions et réponses sont écrites par des annotateurices employé·e·s de l'entreprise, de la manière suivante :

1. Les documents sont découpés en « pages » de 1 000 mots maximum, en s'appuyant sur la structure des titres et sections afin de préserver la cohérence du contenu. Cette taille limite est choisie arbitrairement, avec pour but de limiter la quantité d'information à traiter pour chaque annotation.
2. Les pages obtenues sont distribuées dans un ordre aléatoire aux annotateurices.
3. Chaque annotateurice annote une à une les pages qui lui sont attribuées. L'annotation d'une page se déroule de la manière suivante.
   (a) L'annotateurice écrit une question en rapport avec le contenu de la page, ainsi qu'une réponse correcte à cette question (si la réponse peut être donnée à partir des informations contenues dans la page).
   (b) il/elle sélectionne dans la page un ou plusieurs passages qui permettent de répondre à la question.

   Il est possible de répéter ces étapes afin de produire jusqu'à 5 questions par page.

Il a été demandé aux annotateurices de diversifier, dans la mesure du possible, le type des questions produites, et d'indiquer le type de chaque question au sein d'une liste déroulante. Les types de questions possibles ainsi que leurs nombres sont résumés par la table 1. De plus, 15 questions ne sont associées à aucune réponse, car les informations nécessaires ne sont pas disponibles dans les documents métiers. Dans ce cas il est attendu que le système produise une réponse du type "Je ne sais pas".

Le nombre de passages par question varie de 0 (pour les questions sans réponses) à 4, avec une moyenne de 1,3. On constate que la taille des passages associés aux question varie grandement, allant de 1 mot (l'annotateurice ayant sélectionné un unique mot-clé correspondant à la réponse) à 202 mots. Les implications en terme de choix de métrique d'évaluation sont discutées dans la section 2.3.

2.2 Procédure d'évaluation du système de RAG

Le jeu de questions-réponses à pour vocation de permettre l'évaluation du système de RAG : d'abord les questions sont données en entrée au système, et les sorties du systèmes (réponses générées et passages de document retrouvés) sont collectées ; ensuite les sorties du système sont évaluées grâce à différentes métriques.

La comparaison des réponses générées avec les réponses de référence produit un score global, tandis que la comparaison des passages de documents retrouvés avec les passages de document de référence

Type de question	Nombre	Exemple
booléenne	19	Un client de X qui résilie en raison de l'augmentation de novembre 2022 doit-il payer des frais ?
qui/quoi/où/quand	22	Qui contacter si j'ai un souci avec la solution X ?
comment	17	Comment faire pour consulter les documents et modes opératoires de l'offre X ?
pourquoi	14	Pourquoi le X fourni par Y a-t-il été choisi ?
conditionnelles	17	Que se passe-t-il si X souhaite s'opposer à une cession judiciaire ?
combien	17	Combien de X sont inclus dans Y ?

TABLE 1 – Types de questions présents dans le jeu de questions-réponses. Certaines questions appartiennent à plusieurs types.

produit un score de performance du retriever, qui peut s'avérer utile pour estimer quelle proportion des échecs du système est due au retriever et laquelle est due au générateur, et ainsi cibler les efforts d'amélioration de manière pertinente.

2.3 Métriques disponibles

Cette sous-section décrit brièvement les métriques utilisées pour évaluer le système de RAG. Nous suivons la terminologie de (Ru *et al.*, 2024), où trois types de métriques sont distingués :

— les métriques *de retrieval*, qui évaluent la pertinence des passages retrouvés par rapport à une question donnée ;
— les métriques *globales* (« *overall* »), qui évaluent la réponse générée par le système pour la question ;
— les métriques *de génération*, qui évaluent le comportement du générateur par rapport au contenu des passages issus du retrieval.

Métriques de retrieval. Dans le contexte de la recherche d'information (*information retrieval*), plusieurs métriques sont traditionnellement utilisées afin d'évaluer la qualité du retrieval : rappel, précision, nDCG, MAP, etc. Dans cette étude, nous choisissons de rapporter les scores de rappel, qui seront utilisés comme scores de référence pour évaluer les métriques de retrieval issues des bibliothèques testées. Le *rappel* (*recall*) est défini comme la proportion des éléments pertinents effectivement présents dans les k premiers passages retournés par le retriever. Une approche courante est de calculer le rappel au niveau des documents, c'est à dire en considérant qu'un passage de référence est présent dans les résultats du retriever si au moins un passage issu du même document est présent. Cependant, cette condition n'implique pas que le passage de référence soit effectivement contenu dans un passage retourné ; il est également possible qu'il y ait chevauchement partiel d'un passage de référence et d'un passage retourné. Ce problème se pose en particulier ici, puisque nos références sont faites de passages courts en comparaison des documents dont ils sont issus. Nous calculons donc le rappel au niveau des mots, c'est à dire : le pourcentage de mots des passages de référence présents dans les k premiers passages retrouvés. Notez que les mots sont identifiés par

leur positions dans le texte, et non par leur valeur : des passages retournés totalement disjoints des passages de référence donneront donc un score de 0, même s'ils contiennent des mots identiques. Le choix du rappel par rapport aux autres métriques (telles que la précision) est justifié par le fait que (1) ses scores sont faciles à interpréter, (2) puisque les passages de référence tendent à être relativement réduits, nous cherchons à savoir si ceux-ci sont inclus dans les passages de taille fixe retournés par le retriever, ce qui est une attente raisonnable et correspond à la définition du rappel ; au contraire la précision au niveau des mots donnerait des scores nécessairement bas. Enfin, le choix de cette métrique est soutenu par le fait qu'elle corrèle bien plus aux notes moyennes des évaluateurices ($r = 0.35$) que les autres métriques citées, et notamment que le recall au niveau document ($r = 0.05$).

Les métriques de retrieval proposées par les bibliothèques testées diffèrent des métriques traditionnelles sur deux points principaux : premièrement, certaines d'entre elles se passent de passages de référence, de sorte qu'elles peuvent être appliquées même lorsque ceux-ci ne sont pas disponibles ; deuxièmement, elles font souvent appel à des modèles de langues, ce qui leur permet de calculer le score final en fonction des éléments sémantiquement important des passages considérés. Tout comme le rappel, ces métriques évaluent la qualité du retrieval à partir des k premiers passages retrouvés. Nous choisissons de fixer $k = 5$ pour toutes les métriques, afin que les passages sur lesquels le retrieval est évalué correspondent à ceux effectivement insérés dans le prompt du générateur.

Métriques globales. Les métriques globales évaluent la qualité de la réponse générée selon des critères spécifiques. De nombreuses métriques globales, correspondant à différents critères, sont proposées par les bibliothèques testées. Bien que ces critères soient divers, on peut noter que la plupart s'intéresse à la factualité de la réponse générée, généralement décomposée en deux aspects : la *précision* (« toutes les informations données dans la réponse sont-elles correctes ? »), et le *rappel* (« toutes les informations attendues sont-elles données dans la réponse ? »). À la factualité s'ajoute le critère de *pertinence* (*relevance*) : « toutes les informations données dans la réponses sont-elle en rapport avec la question ? ». En plus des métriques évaluant la factualité et la pertinence, nous intégrons les métriques de *moderation* et de *usefulness* d'Opik : la première vérifie l'absence de propos agressif ou inappropriés, tandis que la seconde mêle divers critères pour obtenir un score général.

Métriques de génération. Les métriques de génération évaluent le comportement du générateur par rapport au contenu des passages issus du retrieval. Le but premier de ces métriques est d'étudier certains comportements du générateur et non sa performance au sens strict ; cependant, on peut vouloir les utiliser pour approximer des critères globaux. Par exemple, la métrique *faithfulness* de DeepEval, qui vise à mesurer la fidélité de la réponse aux passages retrouvés, peut être utilisée pour approximer le critère de précision. Nous intégrons donc à notre étude quelques métriques de génération, qui seront évaluées comme des métriques globales (il s'agit des quatre métriques nommées *hallucination* ou *faithfulness*, voir figure 1).

3 Évaluation des métriques : méthodologie

Nous cherchons à évaluer à quel point les métriques mentionnées dans la section précédente donnent une bonne approximation d'un critère d'évaluation donné. Nous choisissons de définir un critère

global mesurant simultanément des aspects de factualité et de pertinence. Ce critère est évalué sur une échelle de 1 à 5 et défini par le barème de la table 2.

Note	Description
5	Réponse factuelle et pertinente avec la bonne quantité de détails.
4	Réponse factuelle, mais manquant d'informations utiles ou contenant trop d'informations peu importantes.
3	Réponse partiellement correcte, avec de petites erreurs ou approximations OU « je ne sais pas » alors que la réponse de référence contient l'information demandée.
2	Réponse factuellement incorrecte.
1	Réponse hors sujet.

TABLE 2 – Barème utilisé lors de l'évaluation humaine des réponses du système de RAG.

Ce barème est ensuite appliqué par deux évaluateurices (employé·e·s de l'entreprise, ayant une expertise dans le traitement automatique de la langue et les modèles génératifs, dont un a participé à la phase d'annotation pendant la création du dataset), afin de noter chaque sortie du système de RAG sur les 96 instances du jeu de questions-réponses. Nous calculons la corrélation de Pearson [6] entre les deux séries de notes obtenues, afin de vérifier la robustesse du barème et d'estimer la performance maximale pouvant être attendue d'une métrique automatique. La corrélation obtenue est de 0.85. Afin de simplifier les analyses suivantes, nous nous basons sur les notes moyennes obtenues par réponse. Les notes ainsi obtenues sont appelées *notes de références*.

Nous procédons ensuite à l'analyse de corrélation destinée à évaluer les métriques. La section suivant rapporte et analyse :

- les corrélations entre les métriques globales et les notes de référence ;
- les corrélations des métriques de retrieval avec le rappel : bien que le rappel soit lui-même une métrique imparfaite, il est attendu qu'une plus forte corrélation d'une métrique de retrieval avec le rappel dénote une meilleure fiabilité ;
- les corrélations entre les métriques de retrieval et les notes de référence : en effet, puisqu'un meilleur retrieval corrèle avec de meilleures réponses, les métriques de retrieval devraient également corréler avec les notes issues de notre barème.

Toutes les corrélations rapportées correspondent au coefficient de Pearson (effectuer les analyses avec la corrélation de Spearman donne des résultats similaires à ceux rapportés).

4 Résultats

La figure 1 synthétise les corrélations obtenues pour les métriques globales. On remarque tout d'abord que la largeur des intervalles de confiance est substantielle, en raison de la taille modeste de notre échantillon (96). Il est cependant possible de voir certaines tendances apparaître.

Premièrement, on remarque que des métriques parfois jugées obsolètes telles que METEOR corrèlent

6. Nous calculons la corrélation de Pearson plutôt que le kappa de Cohen ou de Fleiss, ceux-ci étant peu adaptés lorsque les annotation sont de nature ordinale, comme c'est le cas ici.

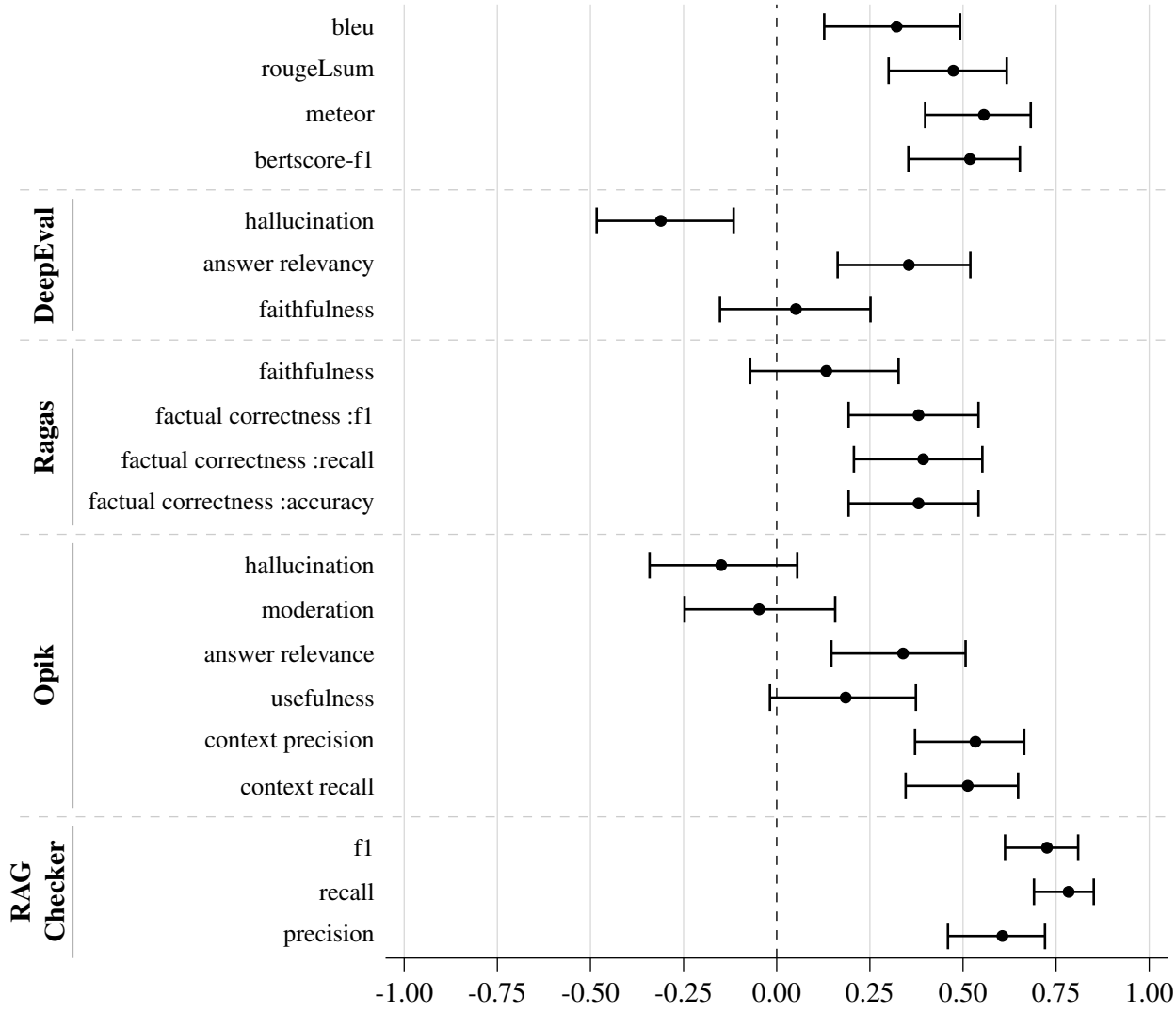

FIGURE 1 – Corrélation des métriques globales avec les notes moyennes des évaluateurices.

étonnamment bien avec les notes de référence. Ensuite, on constate que les métriques de génération utilisées en tant que métriques globales corrèlent faiblement aux notes de références, ce qui est attendu, puisqu'elles n'ont pas accès à la réponse de référence. De même la métrique *moderation* d'Opik ne corrèle pas avec les notes de références. À l'inverse, on observe une très forte corrélation pour les métrique de RAGChecker, en particulier le recall.

La figure 2 résume les corrélations des métriques de retrieval avec le rappel. On constate des chevauchements plus importants des intervalles de confiance. Cependant, il est intéressant de remarquer qu'une métrique telle que la *contextual precision* de DeepEval obtient une corrélation supérieure à 0.5 avec le rappel, et ce sans faire appel aux passages de référence. On observe qu'à l'inverse, les métriques non-LLM de Ragas obtiennent une corrélation très faible. Cela s'explique probablement par le fait qu'elle compare les passages retrouvés aux passages de référence via une distance de Levenshtein, qui ne donne pas de valeur pertinente lorsque les passages comparés sont de tailles très différentes, comme c'est le cas ici.

Si l'on observe la figure 3, on constate un phénomène surprenant : plusieurs métriques montrent une très forte corrélation avec les notes de référence. Cette corrélation est parfois plus forte que la

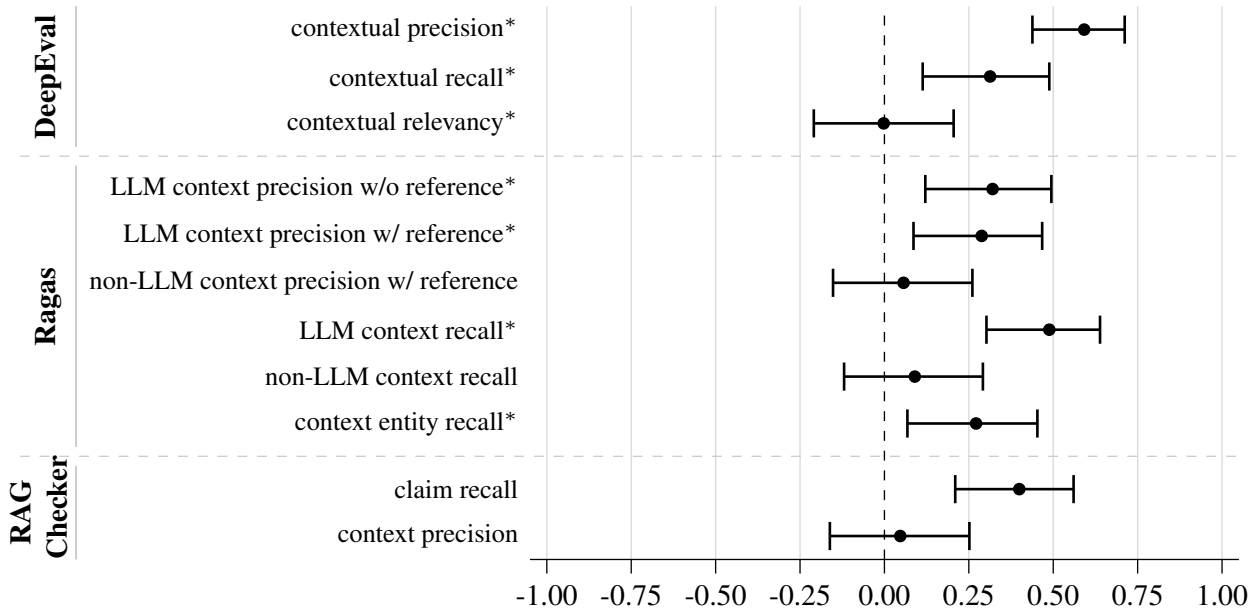

FIGURE 2 – Corrélation des métriques globales avec le rappel. Les métriques ne s'appuyant pas sur les passages de référence sont marquées d'un astérisque.

corrélation avec le rappel. La corrélation du *claim recall* de RAGChecker semble proche de 0,7. Il ne semble pas plausible qu'une telle corrélation soit due uniquement à la capacité de cette métrique de mesurer la pertinence des documents retrouvés. Nous commentons ce phénomène plus en détail dans la section suivante.

5 Limites

La principale limite de notre étude est liée à l'interprétation des corrélations. Il est difficile de savoir a priori sur quoi s'appuient des métriques sophistiquées utilisant des LLM pour produire leurs scores. Par conséquent, observer qu'une métrique donnée corrèle bien avec le jugement humain n'est pas suffisant pour affirmer qu'elle mesure ce que l'on souhaite qu'elle mesure, étant donné la configuration de notre expérience. On peut illustrer cette difficulté en imaginant une métrique mesurant la facilité des questions, sans avoir connaissance des réponses générées. Une telle métrique donnerait des notes plus élevées aux questions faciles, or celles-ci tendent effectivement à obtenir de meilleures notes. Elle aurait donc une corrélation positive avec les notes de référence, alors que sa pertinence pour l'évaluation et la comparaison de systèmes de RAG serait nulle (elle donnerait exactement les mêmes notes à tous les systèmes). Une illustration moins extrême de ce phénomène est sans doute la métrique *claim recall* de RAGChecker, dont la corrélation avec les notes des évaluateurices est très élevée. Il semble probable que cette métrique ne mesure pas uniquement la qualité du retrieval ; une interprétation plausible est qu'elle capture en partie d'autres caractéristiques comme, par exemple, la facilité avec laquelle les informations de la réponse peuvent être extraites des passages retrouvés, ou la facilité de la question. Cette limite est en partie due à la configuration de notre expérience : puisque celle-ci ne comporte qu'un seul système de RAG, il est impossible d'observer le comportement des métriques en fonction des sorties du système indépendamment de la question d'entrée.

Notons que, malgré cette faiblesse, notre méthodologie produit certains résultats exploitables en

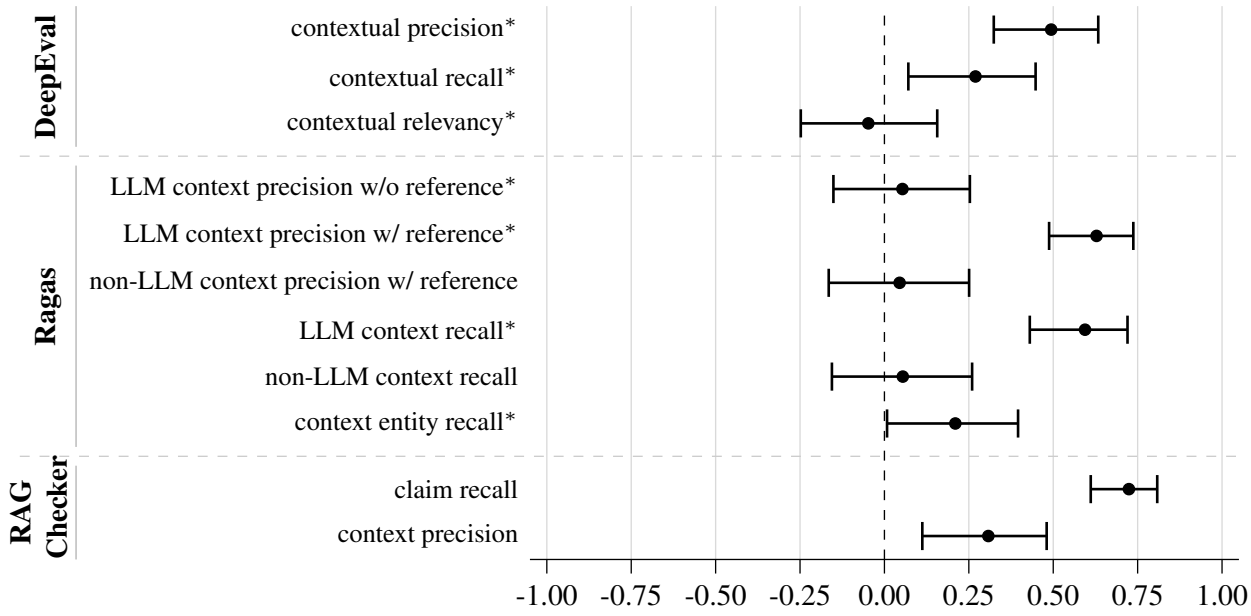

FIGURE 3 – Corrélation des métriques de retrieval avec les notes moyennes des évaluateurices. Les métriques ne s'appuyant pas sur les passages de référence sont marquées d'un astérisque.

éliminant certaines métriques candidates : en effet, on peut affirmer que les métriques ayant une mauvaise corrélation avec le critère d'évaluation considéré ne mesurent pas ce critère.

6 Travaux connexes

Des évaluations de métriques par l'étude de corrélation avec le jugement humain ont été rapportées dans plusieurs publications récentes. Bien que la plupart d'entre elles considèrent l'évaluation de tâches de GLN (Génération du Langage Naturel) autres que le RAG, l'évaluation de ces tâches comporte des enjeux communs à ceux de l'évaluation des réponses d'un système de RAG. Cette section propose un tour d'horizon (non exhaustif) de ces publications, regroupées en fonction de leurs méthodologies d'évaluation.

Tout d'abord, certaines de ces publications s'appuient sur une méthodologie similaire à la nôtre (analyse de corrélation sur un seul système). On peut citer, par exemple : (Liu *et al.*, 2024) qui évalue un système dédié à l'évaluation de divers aspect de diverses tâches de génération du langage naturel, ou (Yeginbergen *et al.*, 2025) qui observe la corrélation de plusieurs systèmes LLM-as-a-judge avec des évaluations humaines de contre-argumentations générées automatiquement. Ces études souffrent de la même limitation méthodologique que la nôtre.

D'autres études évaluent les métriques en mesurant l'adéquation de celles-ci avec un ordre de préférence exprimé sur des paires de sorties correspondant à la même entrée. Plus précisément : chaque entrée de la tâche est associée à deux sorties alternatives, pour lesquelles on dispose d'un ordre de préférence ; toutes les paires pour lesquelles les scores de la métrique respectent l'ordre de préférence sont considérées comme des succès, et le ratio de succès forme score de performance de la métrique. Cette approche permet de se prémunir des facteurs de confusions relatifs aux caractéristiques de la question qui limitent les interprétations de nos résultats. Parmi les études utilisant cette approche,

on peut citer : (Ke *et al.*, 2024; Zhu *et al.*, 2025; Lambert *et al.*, 2025). Certaines études combinent cette approche avec l'analyse de corrélation sur un seul système, par exemple : (Xu *et al.*, 2023; Kim *et al.*, 2024; Xiong *et al.*, 2025).

Enfin, certaines publications rapportent des analyses de corrélation impliquant plusieurs systèmes. Les traitements statistiques effectués peuvent alors varier, puisque les scores générés par une métrique, tout comme les notes de référence, sont alors organisés dans une matrice d'une ligne par système et une colonne par sortie évaluée. Il existe en effet plusieurs manières de calculer une corrélation entre deux matrices : (Gao *et al.*, 2025) étudie quatre d'entre elles, dont deux nous semblent pertinentes dans le cadre de l'évaluation des métriques de RAG. La première est de calculer la corrélation des scores moyens de chaque système noté par la métrique avec la note de référence moyenne de chaque système. La seconde est de calculer, pour chaque entrée, la corrélation scores des métriques avec les notes de référence sur les différents systèmes, puis d'effectuer la moyenne des corrélations ainsi obtenues. Ces deux approches sont également étudiées par (Deutsch *et al.*, 2021). Les deux études concluent empiriquement que la deuxième approche possède un pouvoir plus important de discrimination des métriques testées. De plus, il a été démontré que cette approche réduit considérablement l'importance des facteurs de confusion dans les corrélation mesurées entre métriques d'évaluation de la traduction automatique et le jugement humain (Perrella *et al.*, 2024). Une approche similaire, appliquée dans (Dinh *et al.*, 2024), consiste à calculer la corrélation sur l'ensemble des notes, normalisées par entrées.

7 Pistes de recherche

Les résultats empiriques de notre étude recoupent certains des résultats mentionnés dans la section précédente, et confirment l'utilité d'intégrer plusieurs (au moins deux) systèmes de RAG aux études de corrélation visant à évaluer des métriques de RAG. Il est à noter que les résultats de l'étude seront alors partiellement dépendants des systèmes de RAG choisis : il convient donc de choisir un ensemble de système représentatif de l'ensemble des systèmes auxquels on souhaite appliquer les métriques. De plus, intégrer un trop grand nombre de systèmes risque de rendre la tâche d'annotation coûteuse.

Ces difficultés ouvrent des perspectives de recherches intéressantes. Par exemple, comment estimer l'apport probable d'un modèle ou d'une question supplémentaire à une étude de corrélation ? Est-il possible de développer des procédures pour choisir et adapter le nombre de systèmes et de question à intégrer à l'étude empirique, à mesure que des évaluations humaines sont collectées ? Des réponses à de telles questions pourraient permettre aux professionnels mettant en place des systèmes de RAG de valider de manière plus fiable leurs métriques d'évaluation tout en minimisant les coûts associés à cette validation.

Références

DEUTSCH D., DROR R. & ROTH D. (2021). A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods. *Transactions of the Association for Computational Linguistics*, **9**, 1132–1146. DOI : 10.1162/tacl_a_00417.

DINH T. A., MULLOV C., BÄRMANN L., LI Z., LIU D., REISS S., LEE J., LERZER N., GAO J., PELLER-KONRAD F., RÖDDIGER T., WAIBEL A., ASFOUR T., BEIGL M., STIEFELHAGEN R., DACHSBACHER C., BÖHM K. & NIEHUES J. (2024). SciEx : Benchmarking Large Language

Models on Scientific Exams with Human Expert Grading and Automatic Grading. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 11592–11610, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.647.

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. DOI : 10.18653/v1/2024.eacl-demo.16.

GAO M., HU X., LIN L. & WAN X. (2025). Analyzing and Evaluating Correlation Measures in NLG Meta-Evaluation. In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 2199–2222, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.naacl-long.111.

KE P., WEN B., FENG A., LIU X., LEI X., CHENG J., WANG S., ZENG A., DONG Y., WANG H., TANG J. & HUANG M. (2024). CritiqueLLM : Towards an Informative Critique Generation Model for Evaluation of Large Language Model Generation. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 13034–13054, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.704.

KIM S., SUK J., LONGPRE S., LIN B. Y., SHIN J., WELLECK S., NEUBIG G., LEE M., LEE K. & SEO M. (2024). Prometheus 2 : An Open Source Language Model Specialized in Evaluating Other Language Models. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 4334–4353, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.248.

LAMBERT N., PYATKIN V., MORRISON J., MIRANDA L., LIN B. Y., CHANDU K., DZIRI N., KUMAR S., ZICK T., CHOI Y., SMITH N. A. & HAJISHIRZI H. (2025). RewardBench : Evaluating Reward Models for Language Modeling. In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Findings of the Association for Computational Linguistics : NAACL 2025*, p. 1755–1797, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-naacl.96.

LIU M., SHEN Y., XU Z., CAO Y., CHO E., KUMAR V., GHANADAN R. & HUANG L. (2024). X-Eval : Generalizable Multi-aspect Text Evaluation via Augmented Instruction Tuning with Auxiliary Evaluation Aspects. 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. 8560–8579, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.473.

PERRELLA S., PROIETTI L., SCIRÈ A., BARBA E. & NAVIGLI R. (2024). Guardians of the Machine Translation Meta-Evaluation : Sentinel Metrics Fall In ! In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 16216–16244, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.856.

RU D., QIU L., HU X., ZHANG T., SHI P., CHANG S., JIAYANG C., WANG C., SUN S., LI H., ZHANG Z., WANG B., JIANG J., HE T., WANG Z., LIU P., ZHANG Y. & ZHANG Z. (2024). RAGChecker : A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation. DOI : 10.48550/arXiv.2408.08067.

XIONG T., WANG X., GUO D., YE Q., FAN H., GU Q., HUANG H. & LI C. (2025). LLLaVA-Critic : Learning to Evaluate Multimodal Models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 13618–13628. Conference Name : 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) ISBN : 9798331543648 Place : Nashville, TN, USA, DOI : 10.1109/CVPR52734.2025.01271.

XU W., WANG D., PAN L., SONG Z., FREITAG M., WANG W. & LI L. (2023). INSTRUCTSCORE : Towards Explainable Text Generation Evaluation with Automatic Feedback. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 5967–5994, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.365.

YEGINBERGEN A., ORONOZ M. & AGERRI R. (2025). Dynamic Knowledge Integration for Evidence-Driven Counter-Argument Generation with Large Language Models. Version Number : 2, DOI : 10.48550/ARXIV.2503.05328.

ZHU L., WANG X. & WANG X. (2025). JudgeLM : Fine-tuned Large Language Models are Scalable Judges. arXiv :2310.17631 [cs], DOI : 10.48550/arXiv.2310.17631.