HUB FRANCE IA

GUIDE PRATIQUE

ÉVALUATION DES CHAÎNES DE RAG

Septembre 2025

Table des matières

HUB FRANCE IA

INTRODUCTION

Introduction

L'émergence de l'Intelligence Artificielle Générative (*IAG*) a donné naissance aux chaînes de la génération augmentée par récupération (*RAG*)[1], qui constituent aujourd'hui un levier essentiel pour exploiter efficacement des bases documentaires propriétaires.

Contrairement aux moteurs de recherche classiques qui retournent des documents en fonction de leur similarité lexicale avec la requête, les *RAG* combinent un moteur de recherche capable d'identifier les passages pertinents et un grand modèle de langage (*LLM*)[2] chargé de synthétiser une réponse en langage naturel à partir de ces extraits.

Ce processus repose sur l'intégration d'un *contexte documentaire*[3] directement dans le *prompt* adressé au *LLM*. Ce mécanisme permet ainsi d'exploiter la puissance linguistique des modèles sans recourir à leurs connaissances internes, souvent imprécises ou obsolètes, et s'inscrit comme un pilier des usages maîtrisés de l'IAG.

Les *RAG* sont constitués par une chaîne de traitement incluant à la fois différents pré et/ou post traitements appliqués aux documents, des indexations, des stratégies de recherche et l'usage de *LLMs*. La diversité des traitements et les choix de paramètres et de solutions techniques à chaque étape engendrent une forte combinatoire[4] dans la mise en œuvre d'un *RAG*.

La combinaison de ces éléments au sein de la chaîne rend critique la capacité à évaluer la qualité d'une combinaison donnée et à la comparer à toute autre

---

[1] RAG - *Retrieval Augmented Generation*.

[2] LLM - *Large Language Model*.

[3] Le *contexte documentaire* désigne les extraits de documents, sélectionnés par la phase de recherche d'une chaîne *RAG*, contenant des éléments pertinents pour répondre à une requête utilisateur (le *prompt*). Ces extraits sont alors intégrés au *prompt* envoyé au modèle de langage afin de générer une réponse contextualisée.

[4] Une chaîne RAG n'est pas un composant unique, mais un enchaînement de choix techniques, chacun avec plusieurs variantes possibles. Le nombre de configurations possibles n'augmente pas linéairement : il est multiplicatif (produit du nombre d'options à chaque étape). On parle alors de forte combinatoire.

combinaison. En effet, pour les concepteurs de chaînes de *RAG*, des éléments clés sont souvent manquants pour évaluer la qualité de l'impact d'une nouvelle combinaison sur le système.

Souvent, l'usage d'un *RAG*[5] va de pair avec l'idée d'utiliser le *LLM* uniquement sur des données bien identifiées et pré-indexées, et non sur le savoir global d'Internet. D'une certaine façon, utiliser un *RAG*, c'est une tentative d'utiliser le modèle de langue pour ses capacités linguistiques et d'analyses, et non pour ses connaissances.

Ce document a pour objectif d'apporter au lecteur un cadre méthodologique et des outils pour évaluer la qualité de sa chaîne de *RAG*. Plus précisément, nous aborderons les points suivants :

- La méthodologie.
- Les jeux de données :
    - Définitions des prérequis pour la construction d'un jeu de données de test (*Ground Truth*).
    - Sélection de jeux de données existants utilisables dans diverses situations.
- Les outils d'évaluation : identification des principales solutions disponibles pour l'évaluation des chaînes de *RAG*, tant open-source que commerciales.
- La définition des métriques standards pour l'évaluation de la qualité des chaînes de *RAG*.
- Quelques résultats obtenus lors de nos tests.

[5] Consulter la fiche pratique : Hub France IA. Fonctionnement d'un *RAG*. Groupe de travail IA Générative – Bonnes pratiques. Mars 2025. https://www.hub-franceia.fr/hub-france-ia-fiche-pratique-ia-generative-comment-fonctionne-un-rag/

HUB FRANCE IA

1. Audience

1. Audience

L'évaluation des chaînes de *RAG*, visant à déterminer la qualité de restitution des résultats des *RAG* et la pertinence des informations générées par l'IA, intéressera principalement :

- Les ***data scientists*** chargés de proposer et mettre en œuvre un *RAG*. Ils trouveront dans cette évaluation des informations tangibles pour guider leurs choix.
- Les **développeurs** qui disposeront d'un moyen de quantifier l'amélioration ou la dégradation des résultats en fonction des modifications apportées au système.
- Des **décideurs** (DSI, RSSI) avant de valider l'adoption d'une solution de *RAG* au sein de leur entreprise en s'appuyant sur les résultats d'évaluations méthodiques.

Ce document s'adresse à toute personne impliquée dans la définition, la construction ou l'évaluation des chaînes de *RAG*. Bien qu'il ne soit pas nécessaire d'avoir des compétences en développement pour comprendre et utiliser ce document, celles-ci sont souvent requises pour mettre en place un mécanisme d'évaluation. En effet, comme exposé dans ce document, les outils d'évaluation et de mise en place des chaînes de *RAG* sont souvent distincts, et un travail d'intégration ou d'ajustement du format des données est nécessaire.

HUB FRANCE IA

2. Contribution

2. Contribution

L'un des objectifs principaux du groupe de travail était de fournir à nos lecteurs une vision globale de la problématique de l'évaluation des chaînes de *RAG*. À cette fin, le groupe de travail a combiné des éléments scientifiques relatifs aux métriques et à la mesurabilité avec des aspects pratiques tels que les outils et les jeux de données facilement accessibles et utilisables dans des contextes industriels ou académiques. La couverture de tous ces aspects dans un guide pratique et accessible constituait le premier objectif du groupe de travail et représente sa principale contribution.

Dans le cadre de notre veille scientifique, nous avons fréquemment observé des évaluations se concentrant sur les différentes parties des chaînes de *RAG,* soit sur la partie *Retrieval*, soit sur la partie Générative, mais moins de réflexions approfondies sur la combinaison des deux, et en particulier sur l'utilisation et l'évaluation des « Réponses avec Citations » en sortie des chaînes de *RAG*. On trouve des travaux sur la précision des citations[6] et nous avons souhaité creuser les métriques sur le sujet[7]. L'utilisation des citations nous semble être un élément crucial pour garantir la véracité des réponses et constitue une mesure de protection efficace contre les hallucinations. Nous avons approfondi notre réflexion sur le postulat suivant : « Une bonne réponse est une réponse correcte, argumentée et avec les sources citées ». Ce postulat est-il toujours d'actualité ? Nous avons commencé à explorer des pistes de métriques allant dans ce sens.

---

[6] Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E. Ho, James Zou. How well do LLMs cite relevant medical references? An evaluation framework and analyses. February 2024. *arXiv preprint arXiv: 2402.02008.* https://arxiv.org/abs/2402.02008 et Simon Knollmeyer, Oğuz Caymaze, Leonid Koval, Muhammad Uzair Akmal, Saara Asif, Selvine G. Mathias, Daniel Großmann. Benchmarking of Retrieval Augmented Generation: A Comprehensive Systematic Literature Review on Evaluation Dimensions, Evaluation Metrics and Datasets. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024)-Volume 3*. SciTePress, 2024. p. 137-148. https://www.scitepress.org/Papers/2024/130657/130657.pdf

[7] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, Zhaofeng Liu. Evaluation of Retrieval-Augmented Generation: A Survey. In *CCF Conference on Big Data*. Singapore. Springer Nature Singapore, 2024. p. 102-120. https://arxiv.org/abs/2405.07437v2

Dans ce contexte, nous proposons également un « format » pour les jeux de tests, qui inclut non seulement la question et la réponse attendue, mais également les documents requis. Ce « format » nous permet d'opérer et de proposer une nouvelle métrique d'évaluation globale de la qualité d'une chaîne de *RAG*.

Enfin, conformément aux éléments précédents et comme évoqué en introduction, l'un des principes du *RAG* est d'utiliser un modèle d'IAG en combinaison avec un corpus documentaire distinct. Il s'agit ainsi de tirer parti des **capacités linguistiques du modèle**, et non de ses **connaissances intrinsèques**[8].

[8] Ce point, qui fait également l'objet de notre réflexion, est abordé plus en détail dans : Nathan Atox, Mason Clark. LLMs. Evaluating Large Language Models through the Lens of Linguistic Proficiency and World Knowledge: A Comparative Study. *Authorea Preprints*. August 2024. https://www.authorea.com/doi/full/10.22541/au.172479372.22580887

HUB FRANCE IA

3. Méthodologie

3. Méthodologie

Pour aborder le sujet en profondeur, nous présentons une vue d'ensemble des composants impliqués et de leurs interrelations.

3.1. Le jeu de test

Le jeu de données de test est un prérequis. Pour évaluer un système de *RAG*, il faut se doter 1) d'un ensemble de documents (le *contexte documentaire*) qui seront indexés et fourniront le socle de la partie « *Retrieval* », et 2) de paires de questions/réponses avec des réponses vérifiées par un humain. L'évaluation consistera à vérifier si les réponses générées automatiquement en sortie de la chaîne de *RAG* sont en ligne avec la réponse correcte de l'humain.

Voir la figure 1 « Composition des jeux de test » comme exemple d'un mini jeu de tests, contenant X documents et un ensemble de questions / réponses :

- Document : X documents, par exemple un document *Les grandes tours.pdf*, contenant le texte « La 101 Tower, TAIPEI, 101 étages, 508m »
- Question : quelle est la plus haute tour à Taipei ?
- Réponse attendue : la plus haute tour à Taipei est la Taipei 101 avec ses 508m de hauteur.

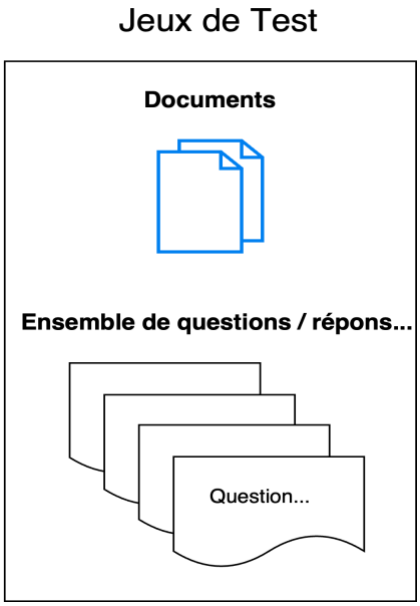

Figure 1 : Composition des jeux de

3.2. Indexation des contenus

Les documents doivent être injectés dans un moteur de recherche. Cette étape est généralement réalisée en amont (et non en temps réel), via une mécanique d'indexation. Il existe de nombreuses stratégies de moteurs de recherche :

Vector Store

Vectorisation de contenus, indexation, recherche par similarité vectorielle, score Cosinus[9].

Principe : les documents sont transformés en vecteurs de haute dimension via des modèles d'*embedding* (*BERT, Sentence-BERT, OpenAI Ada*, etc.). La recherche s'effectue par calcul de similarité cosinus entre le vecteur de la requête et ceux des documents indexés.

Avantages : prise en compte de la sémantique, gestion efficace des synonymes et concepts sémantiquement proches.

Inconvénients : coût de calcul élevé. Mauvaises performances pour des recherches de mots très précis comme des codes d'erreurs ou des références.

Mots Clés

Tokenisation des contenus, indexation de *tokens* (mots clés). Recherche par *token*. Scores *BM25*[10]*, TF-IDF*, etc.

---

[9] Harald Steck, Chaitanya Ekanadham et Nathan Kallus. Is Cosine-Similarity of Embeddings Really About Similarity? In *Companion Proceedings of the ACM Web Conference 2024*, p. 887-890. https://arxiv.org/abs/2403.05440 Les auteurs questionnent si la similarité cosinus des *embeddings* reflète vraiment la similarité sémantique. L'étude met en évidence des biais et limite la fiabilité des mesures cosinus dans certains contextes, pointant la nécessité d'analyses plus fines dans les systèmes de recherche vectorielle.

[10] Une synthèse classique sur BM25 et ses fondements théoriques et probabilistes : Stephen Robertson et Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. In *Foundations and Trends in Information Retrieval*. Vol. 3, No. 4, 2009, p 333–389. https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf

Principe : les documents sont découpés en *tokens* qui sont indexés dans une structure inversée. La recherche utilise des algorithmes de *scoring* comme *BM25* pour classer les résultats selon la pertinence lexicale.

Avantages : très rapide, excellent pour les correspondances exactes, mature et scalable.

Inconvénients : limité par la correspondance lexicale, difficultés avec les synonymes.

Sémantique et Graphe[11]

Indexation basée sur les relations entre entités. Recherche par « *Graph Traversal* ».

Principe : les documents et leurs métadonnées sont modélisés comme un graphe d'entités reliées. La recherche exploite non seulement les relations sémantiques entre concepts, mais aussi les relations logiques ou non-sémantiques.

L'intérêt du graphe est précisément de capturer des éléments factuels et relationnels qui échappent à la représentation vectorielle, celle-ci étant purement sémantique.

Avantages : excellent pour les requêtes complexes impliquant des relations, découverte de connexions implicites.

Inconvénients : complexité de modélisation, performance variable selon la structure du graphe.

Approches hybrides

Combinaison de plusieurs stratégies pour optimiser précision et rappel selon les cas d'usage.

---

[11] Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, Wei Hu. Knowledge Graph-Guided Retrieval Augmented Generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1 Long Papers, p. 8912–8924, 2025. https://aclanthology.org/2025.naacl-long.449/

Principe : combinaison de plusieurs stratégies d'indexation — par exemple un moteur lexical (*BM25, TF-IDF*) et un moteur vectoriel (*embeddings* + similarité cosinus) — afin de tirer parti des forces de chacune. La recherche peut être effectuée en parallèle dans plusieurs index puis fusionnée (pondération des scores, *Reciprocal Rank Fusion*) ou traitée de manière séquentielle (filtrage lexical suivi d'un affinage vectoriel, ou inversement)[12].

Avantages

- Complémentarité des signaux, amélioration simultanée du rappel et de la précision, résilience aux variations de formulation des requêtes, et, flexibilité pour ajuster la pondération selon le domaine ou le cas d'usage.

Inconvénients

- Complexité de mise en œuvre accrue, coût de calcul et latence plus élevés en raison de recherches multiples et du *re-ranking*, optimisation plus difficile du fait de l'interaction entre les paramètres des différentes stratégies, et, dépendance à la qualité des composants : un mauvais réglage d'une partie peut dégrader l'ensemble.

3.3. Pré-traitements

Il est important de souligner que, en amont de l'indexation, un travail de prétraitement des documents peut être nécessaire. En effet, la chaîne de traitement comprend souvent des étapes telles que :

- Extraction des données : par exemple, extraction du texte depuis un *PDF*, une image, une vidéo.
- Conversion des données : par exemple conversion d'*HTML* en texte.
- Optimisation des données : par exemple, correction automatique, normalisation des termes, etc.

[12] Dohyeon Lee, Seung-won Hwang, Kyungjae Lee, Seungtaek Choi, Sunghyun Park. On Complementarity Objectives for Hybrid Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Vol. 1 Long Papers, p. 13357–13368. July 9-14, 2023. Questionne l'intérêt des approches hybrides entre retrievers denses et clairsemés.

- Découpage des données : découpage d'un document en entités documentaires plus petites, par page, paragraphe, nombre de mots, ou toute autre stratégie de « chunking ».

Voici, par exemple, la figure 2 « Exemple d'une chaîne d'indexation dans un *Vector Store* » illustrant une chaîne d'indexation classique pour une indexation dans un *Vector Store*. Les éléments qui constituent la chaîne de traitement, soit la « *pipeline* », peuvent bien sûr différer d'une implémentation à l'autre et chaque acteur va se distinguer par la qualité de sa « *pipeline* » de traitement.

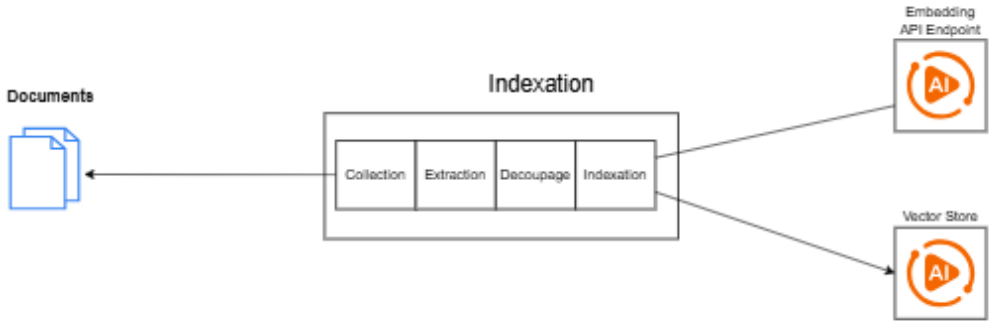

Figure 2 : Exemple d'une chaîne d'indexation dans un *Vector Store*

Il est déjà intéressant de constater que de nombreux éléments de la chaîne de traitement sont déjà pris en compte. Chacun d'entre eux pourra être considéré comme un facteur de variabilité susceptible d'influencer le résultat final.

3.4. Réponse à une question

In fine, une chaîne de *RAG* utilise un moteur de recherche associé à un modèle d'IAG afin d'obtenir une réponse à une question.

Dans le cas classique d'une chaîne de *RAG* avec un *Vector Store* :

- La question va être vectorisée.
- Le Vector Store sélectionne les documents le plus similaires à la question (recherche sémantique par similarité Cosinus).
- La question et les documents sélectionnés sont combinés via un « *prompt* textuel » et envoyés au modèle d'IAG.

Il est possible d'effectuer des post-traitements entre l'obtention des résultats de la recherche et la construction du *prompt* textuel. Un exemple classique est

l'utilisation d'un modèle de réordonnancement qui évalue la pertinence du contenu de chaque document par rapport à la question posée et puis les trie par ordre de pertinence décroissante. L'ajout d'une telle étape introduit des paramètres supplémentaires à évaluer dans la chaîne de *RAG* (nombre de documents à fournir au modèle et nombre de documents à conserver en sortie) et peut avoir un impact significatif sur la qualité des réponses finales, les *LLM* exploitant généralement mieux le début ou la fin de leur contexte.

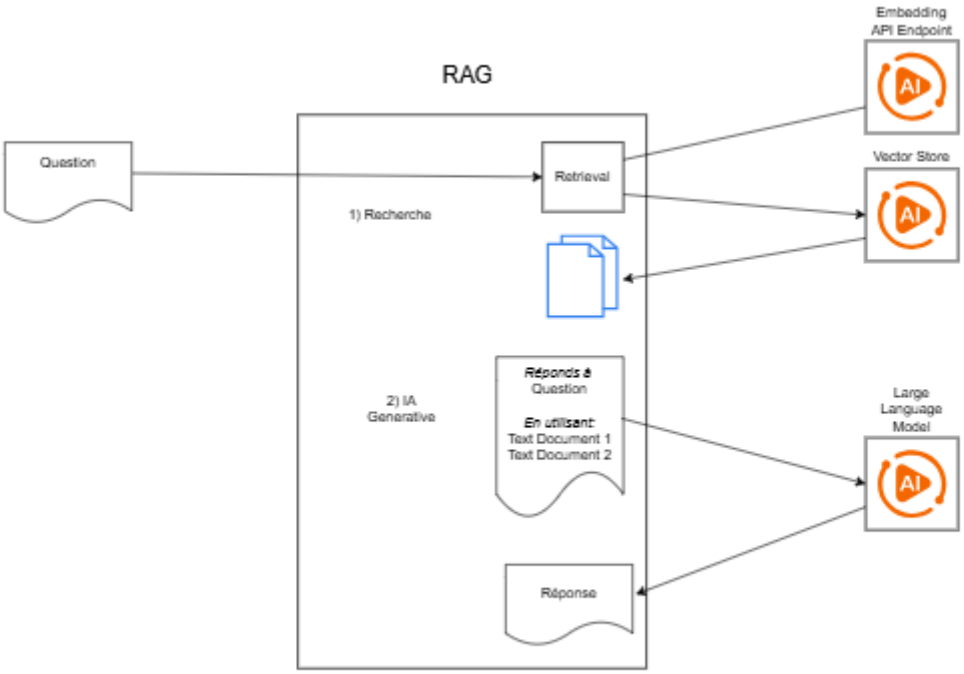

Figure 3 : Éléments constitutifs d'une chaîne de *RAG* avec *Vector Store*

Comme illustré dans l'exemple de la figure 3 « éléments constitutifs d'une chaîne de *RAG* avec *Vector Store* », le point d'entrée est une question (le *prompt*), la sortie est une réponse. C'est cette réponse que nous allons vouloir comparer avec la « Réponse attendue » du jeu de test.

Ainsi, si nous reprenons le jeu de données et la partie « *RAG* », l'évaluation portera principalement sur la sortie du *RAG*, en comparant la « Réponse attendue », décrite dans le jeu de test, avec la « Réponse obtenue », en sortie de la chaîne de *RAG*.

HUB FRANCE IA

4. Évaluation d'un système de RAG

4. Évaluation d'un système de *RAG*

L'évaluation des chaînes de *RAG* est une étape essentielle pour garantir la qualité, la pertinence et la fiabilité des réponses produites. Globalement, l'évaluation consiste à comparer les résultats issus du système avec les résultats tels que fournis par un humain.

L'évaluation ne se limite pas à mesurer la justesse des réponses retournées par le système, mais examine aussi chacune des étapes clés de la « *pipeline* ». Dans cette section, nous détaillons d'abord l'**évaluation de la recherche documentaire** pour juger de la pertinence et de la couverture des documents retrouvés. Nous abordons ensuite l'**évaluation des citations** qui vérifie la cohérence entre la réponse et ses sources. Puis nous combinons ces dimensions dans l'**évaluation conjointe des réponses et des citations** pour une vision holistique. Nous évoquons également les points de variation et pistes d'optimisation afin d'identifier les leviers d'amélioration de la chaîne de *RAG*. Enfin, la section Métriques de ce document abordera les différentes options permettant d'évaluer la similarité sémantique entre les deux réponses, cette dernière constituant la principale métrique d'évaluation qui fait consensus. Nous présentons les principales métriques utilisées pour quantifier ces performances et comparer différentes configurations. Étant donné que les réponses attendues et obtenues sont textuelles, la notion d'équivalence se révèle complexe.

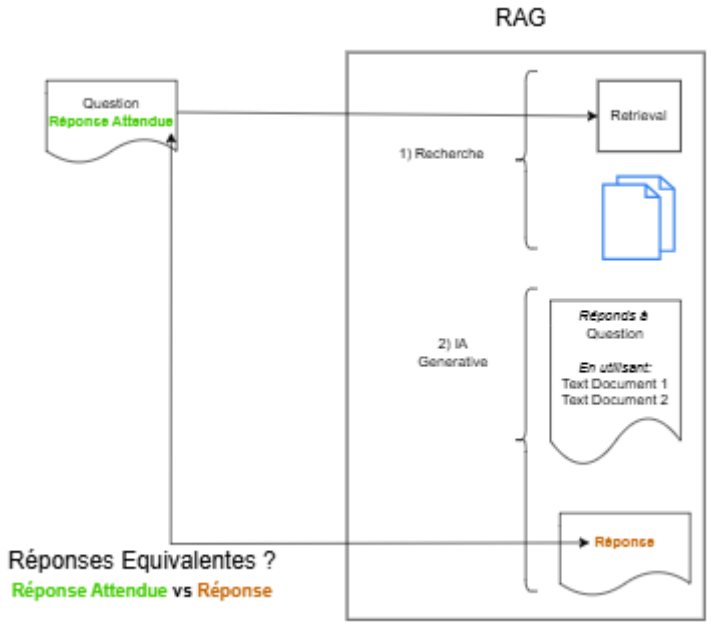

Figure 4 : Éléments constitutifs de l'évaluation simple d'une chaîne de *RAG*

4.1. Évaluation de la recherche documentaire

Comme illustré précédemment, la chaîne de *RAG* se divise en deux phases distinctes : la phase de « *Retrieval* » ou « Recherche » et la phase de génération de texte par l'IA.

L'approche traditionnelle consiste à évaluer la phase de « recherche documentaire » de façon indépendante. En effet, une recherche documentaire efficace conditionne fortement les étapes suivantes : si les documents pertinents ne sont pas identifiés, le contexte fourni au *LLM* ne contiendra pas suffisamment d'éléments factuels pour répondre à la requête. Cela engendre un risque élevé d'hallucinations, de paraphrases erronées ou de réponses factuellement incorrectes de la part du *LLM*.

Une solution consiste donc à scinder l'évaluation en deux étapes :

- Évaluation de la recherche documentaire.
- Évaluation de la génération de réponse.

L'évaluation de la recherche documentaire nécessite un jeu de données de test spécifique comme illustré dans les figures 5 et 6 « compositions d'un jeu de tests », comprenant :

- Des phrases de test.
- Des articles attendus.

De ce modèle de test nous pourrons alors générer les métriques classiques de la recherche documentaire, tel que les *F Score* qui seront décrits dans le chapitre 6.

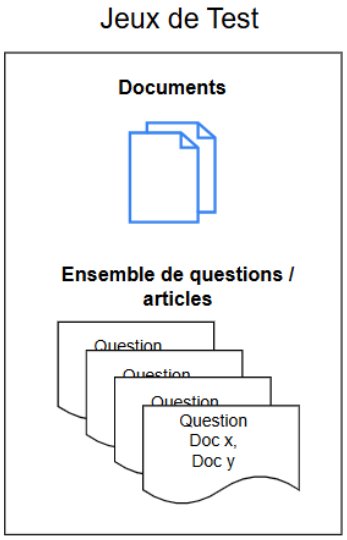

Figure 5 : Composition d'un jeu de tests

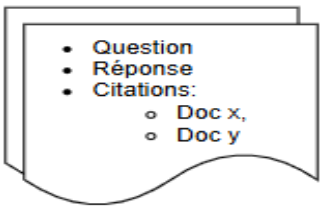

Figure 6 : Composition d'un jeu de tests

4.2. Évaluation des citations

L'un des objectifs principaux d'une chaîne de *RAG* est souvent d'obtenir non seulement une réponse de qualité, mais également d'assurer l'explicabilité de cette réponse. Cela passe par l'utilisation de citations, qui permettent de « prouver » l'origine des assertions et limitent considérablement le risque d'hallucination. Les citations sont essentielles dans des contextes critiques (exemple : juridique, médical, scientifique).

Si le *prompt* demande la citation des sources, ce qui est souvent un prérequis à l'utilisation de chaînes de *RAG* dans un contexte industriel, il est envisageable de combiner les deux types de jeux de tests.

Ainsi, nous pourrions concevoir une métrique combinant la précision des réponses et l'exactitude des sources citées.

4.3. Évaluation réponse et citations

En combinant la partie « Réponse » et la partie « Citations », nous pouvons ainsi évaluer un grand nombre de points de la chaîne *RAG* (voire figure 7 « Éléments constitutifs de l'évaluation double d'une chaîne de *RAG* ») :

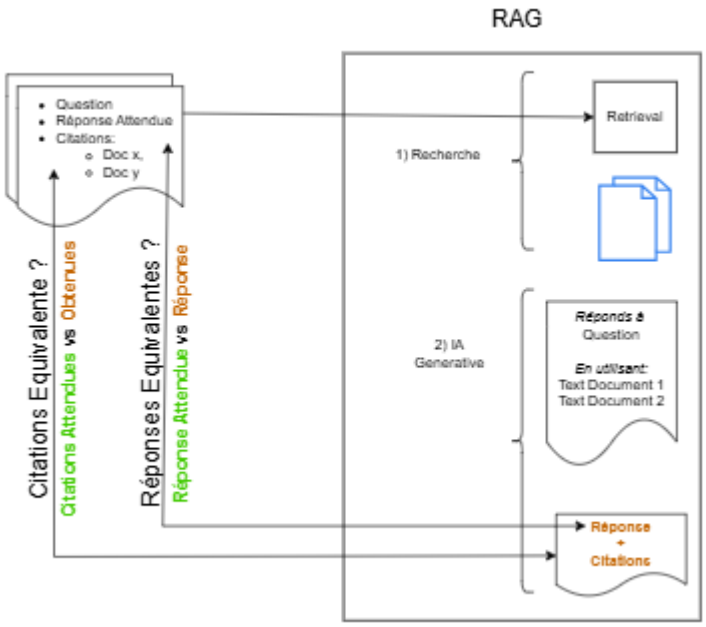

Figure 5 : Éléments constitutifs de l'évaluation double d'une chaîne de *RAG*

Il est important de noter que dans l'exemple précédent, seul le résultat final du *RAG* est présenté. Afin d'optimiser ce résultat, il est nécessaire que :

- Le « *Retrieval* » retourne les articles pertinents.
- Le modèle d'IAG génère une réponse adéquate en s'appuyant sur les articles fournis et la question posée.
- Le modèle d'IAG identifie et cite correctement les sources utilisées pour construire sa réponse.

Un biais de cette méthode est que l'analyse d'un résultat insatisfaisant rendra difficile la détermination de la source principale du problème, qu'elle provienne de la partie Retriever ou de la partie Générative, comme illustré dans la figure 8 « Éléments constitutifs de l'évaluation triple d'une chaîne de RAG ». C'est pourquoi nous proposons d'inclure un troisième point d'évaluation, spécifiquement dédié à la partie *Retrieval*.

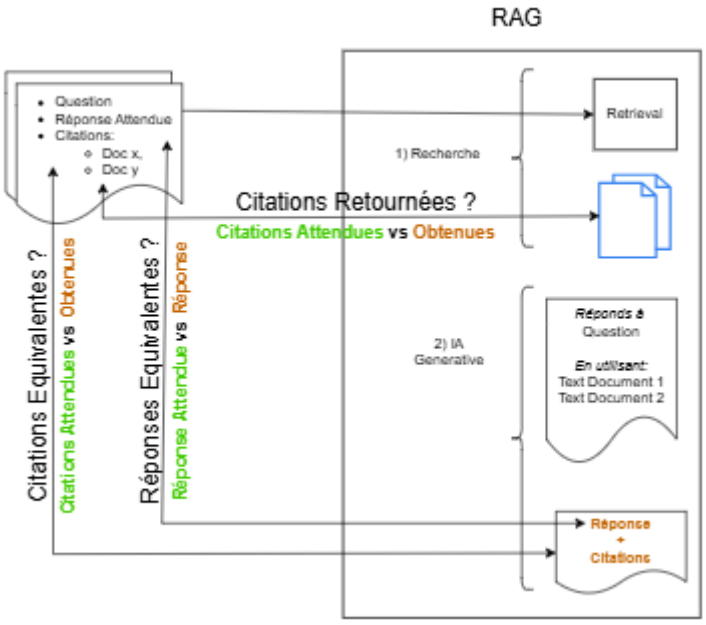

Figure 6 : Éléments constitutifs de l'évaluation triple d'une chaîne de RAG

Il est aussi important de noter que la capacité (partielle) du modèle d'IAG à ignorer les éléments de contexte non pertinents rend le Rappel plus important que la Précision dans le processus de *Retrieval*. En effet :

- L'impact d'un faux positif sera faible, car le modèle d'IA Générative est généralement capable de l'ignorer.
- L'impact d'un article manquant sera par contre important, car le modèle d'IA Générative ne disposera pas de tous les éléments nécessaires pour répondre à la question.

Cette méthode inclut donc trois points d'évaluation :

- Évaluation du Retrieval avec le taux d'inclusion des citations attendues.
- Évaluation du modèle d'IAG : qualité de la réponse.
- Évaluation du modèle d'IAG : couverture des citations.

Une proposition de formule pour cette métrique sera présentée dans la section Métriques.

4.4. Les points de variation et l'optimisation

Comme évoqué précédemment, la chaîne de *RAG* implique de nombreuses étapes, chacune associée à des paramètres qui influencent, individuellement et collectivement, la qualité des réponses.

Ainsi, le bloc « Indexation » présente plusieurs points de variation (et donc d'évaluation) possibles. Par exemple :

Pré-traitement appliqué aux documents :

- Format.
- Extraction.
- Chunking : modèle de découpage, taille des chunks.

Indexation des documents avec le choix des moteurs :

- D'un moteur d'indexation par mots clés (BM25).
- D'un modèle de vectorisation (applicable dans le cadre d'un Vector Store).D'un autre type de moteur (*Graphe*, règles *NLPs*, etc.)

Le bloc « **Évaluation** » dans lequel on constate différents points de variation :

- Pré-traitement sur la phrase utilisateur.
- Éventuel modèle de vectorisation de la phrase utilisateur.
- Recherche des documents correspondants au nombre de documents renvoyés.
- Construction du *prompt*.
- Génération de la réponse A avec un LLM sélectionné.

Ainsi, une fois une bonne mécanique d'évaluation en place, on peut imaginer faire des évaluations sur différentes variations de ces paramètres, pour ensuite les traiter :

- Individuellement : faire varier un paramètre, relancer l'évaluation et observer l'impact de ce paramètre sur le résultat global.
- Collectivement : il n'est pas certain que les paramètres puissent être optimisés individuellement. L'utilisation d'un paramètre optimisé individuellement ne garantit aucunement un résultat optimal. Une meilleure combinaison pour l'ensemble peut exister.

HUB FRANCE IA

5. Jeux de données

5. Jeux de données

Il est relativement aisé de trouver des jeux de données applicables aux problématiques de recherche documentaire. En revanche, il est plus complexe de trouver des ensembles de données contenant les éléments nécessaires à l'évaluation des chaînes de *RAG*, à savoir :

- Documents.
- Questions.
- Réponses.
- Citations (optionnel).

Les sections suivantes présentent en détail les jeux de données sélectionnés pour nos expérimentations. Contrairement aux jeux de données classiques du domaine des systèmes de questions-réponses (tels que *SQuAD, FQuAD, TREC* ou *Natural Questions*), largement conçus pour l'extraction de réponses factuelles dans des contextes bien délimités, notre sélection privilégie des jeux adaptés à l'évaluation de systèmes *RAG* en contexte ouvert et thématique. Ces jeux ont été retenus pour leur capacité à simuler des cas d'usage réalistes, intégrant à la fois diversité des sources documentaires et complexité des interactions. Un aperçu synthétique des jeux retenus, incluant leur licence, leur volumétrie et leur pertinence pour nos tests, est présenté dans le Tableau 1 : Tableau récapitulatif des jeux de données.

Jeux de données	**License**	**Documents**	**Questions/ Réponses**	**Citation (*Context*)**
Single Topic RAG Evaluation	*MIT*	20	120	Non
Acquired Podcast Transcript	*Creative Common* 1.0	200	80	Oui

Tableau 1 : Tableau Récapitulatif des Jeux de Données.

5.1. Single-Topic RAG Evaluation Dataset

L'un des jeux de données utilisés est intitulé « *Single-Topic RAG Evaluation* », créé par Samuel Matsuo Harris et publié sous licence *MIT* sur la plateforme *Kaggle*[13]. Il est composé de quatre fichiers au format CSV, chacun contenant des textes, des questions associées et les réponses correspondantes.

Les fichiers présentent différentes configurations :

- Un fichier contient uniquement les textes et leurs liens sources d'origine.
- Un autre fichier inclut des questions avec des réponses dispersées dans plusieurs extraits du texte associé.
- Un troisième fichier propose des questions pour lesquelles les textes ne contiennent pas de réponse pertinente.
- Enfin, un quatrième fichier contient des questions dont la réponse est présente dans un seul extrait du texte (et non répartie à travers plusieurs parties).

Il convient de noter que ce jeu de données contient uniquement 20 textes distincts, chacun étant accompagné de 6 questions réparties équitablement selon trois catégories :

- 2 questions avec des réponses s'appuyant sur un seul extrait du texte.
- 2 questions avec des réponses issues de plusieurs extraits.
- 2 questions pour lesquelles aucune réponse n'est disponible dans le texte.

5.2. Acquired Podcast Transcripts

Un autre jeu de données pertinent, intitulé « *Acquired Podcast Transcripts and RAG Evaluation* », a été créé par Harry Wang et son équipe (avec des contributions de Rain Jiang, Yihong (Eric) Chen et des étudiants de son cours d'Intelligence Artificielle Générative du printemps 2024), et publié aussi sur la

[13] Samuel Matsuo Harris. *Single-Topic RAG Evaluation Dataset.* https://www.kaggle.com/datasets/samuelmatsuoharris/single-topic-rag-evaluation-dataset

plateforme Kaggle [14]. Cet ensemble de données comprend les éléments principaux suivants :

- **Transcriptions de podcasts** : il contient 200 transcriptions du podcast *Acquired*, collectées sur le site officiel (*acquired.fm*). Les fichiers de transcription individuels sont fournis, principalement au format .txt.
- **Métadonnées des transcriptions** : un fichier acquired_metadata.csv accompagne les transcriptions, spécifiant diverses métadonnées associées.
- **Ensemble de données Questions-Réponses (QA) pour l'évaluation RAG** : un fichier acquired-qa-evaluation.csv a été développé spécifiquement pour évaluer les systèmes RAG. Ce fichier contient 80 ensembles uniques de questions-réponses et comprend les colonnes suivantes :
    - ***Question*** : la question posée pour l'évaluation.
    - ***Human_answer*** : la réponse à la question, fournie par un humain.
    - ***AI_answer_without_the_transcript*** : la réponse générée par un modèle d'IA n'ayant pas eu accès à la transcription concernée.
    - ***AI_answer_without_the_transcript_correctness*** : l'évaluation de l'exactitude factuelle (vérifiée par un humain) de la réponse de l'IA formulée sans accès à la transcription.
    - ***AI_answer_with_the_transcript*** : la réponse générée par un modèle d'IA ayant eu accès à la transcription concernée.
    - ***AI_answer_with_the_transcript_correctness*** : l'évaluation de l'exactitude factuelle (vérifiée par un humain) de la réponse de l'IA formulée avec accès à la transcription.
    - ***Quality_rating_for_answer_with_transcript*** : une note de qualité (attribuée par un humain) pour la réponse de l'IA ayant eu accès à la transcription.
    - ***Post_url*** : l'URL de l'épisode de podcast correspondant à la question.
    - ***File_name*** : le nom du fichier de transcription associé à l'épisode.

[14] Harry Wang. Acquired Podcast Transcripts and RAG Evaluation. https://www.kaggle.com/datasets/harrywang/acquired-podcast-transcripts-and-rag-evaluation

5.3. *RAGProbe* – Auto-génération d'un jeu de tests

Une autre approche consiste à générer les jeux de tests dynamiquement, sur la base du jeu documentaire. Cette approche est décrite précisément par *RAGProbe*[15]. Un des gros avantages de cette approche est qu'elle peut être automatisée complètement et donc être intégrée directement dans une chaîne CI/CD, faisant ainsi partie intégrante du processus : directement après l'indexation, de n'importe quel corpus documentaire, il serait immédiatement possible d'avoir une métrique de qualité de la chaîne de *RAG*.

[15] Shangeetha Sivasothy, Scott Barnett, Stefanus Kurniawan, Zafaryab Rasool, Rajesh Vasa. RAGProbe: An Automated Approach for Evaluating RAG Applications. *arXiv preprint arXiv:2409.19019.* 2024. https://arxiv.org/abs/2409.19019

HUB FRANCE IA

6. Métriques

6. Métriques

L'évaluation des systèmes *RAG* repose sur l'utilisation de métriques. Comme évoqué précédemment, il est possible (mais non obligatoire) de scinder le problème en deux aspects :

- Recherche Documentaire
- Génération de Réponse

Ainsi, les métriques peuvent être considérées sous l'un de ces deux angles, ou de manière plus globale en tentant de les combiner.

6.1. Recherche documentaire

La problématique de l'évaluation de la recherche documentaire est ancienne et a fait l'objet de nombreuses études. Les métriques communément associées sont le **Rappel**, la **Précision**, le ***F-Score***.

Ces métriques évaluent les rapports entre les Documents Pertinents et l'ensemble des Documents. Il est important de noter que ces métriques ont été conçues en fonction des besoins de la recherche documentaire de l'époque, notamment les moteurs de recherche. Ces derniers étaient généralement confrontés à un besoin de rappel important, comme le démontre constamment Google avec ses millions de résultats. Dans le contexte du *RAG*, avec une fenêtre de contexte limité (limite technique et de coût), il est primordial de maximiser la Précision.

Si la Précision et le Rappel restent les principales métriques d'évaluation de la recherche documentaire, les auteurs de *Evaluation of Retrieval-Augmented Generation: A Survey*[16] proposent des métriques alternatives :

- **Relevance ou pertinence** (Relevant Documents ↔ Query) : évalue la correspondance entre les documents récupérés et les informations requises

[16] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, Zhaofeng Liu. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*. Singapore: Springer Nature Singapore, 2024. p. 102-120. https://arxiv.org/pdf/2405.07437

dans la requête. Elle mesure la précision et la spécificité du processus de recherche.

- ***Accuracy* ou exactitude** (*Relevant Documents* ↔ Documents Candidates) : évalue l'équivalence des documents récupérés par rapport à un ensemble de documents candidats. Il s'agit d'une mesure de la capacité du système à identifier les documents pertinents et à leur attribuer une note supérieure aux documents moins pertinents ou non pertinents.
- **Rappel ou sensibilité** : mesure la proportion de documents pertinents retrouvés par rapport à l'ensemble des documents pertinents disponibles dans la base. C'est une mesure de couverture : est-ce que le système récupère tout ce qui pourrait être utile ?
- **Précision** : c'est le taux de documents pertinents parmi l'ensemble des documents proposés.
- **Précision@k** : c'est le taux de documents pertinents parmi l'ensemble des documents proposés dans la liste des k documents récupérés en tête. Elle mesure l'exactitude de ces k documents.
- **Rappel@k** : c'est le taux de documents pertinents correctement identifiés dans les k premiers documents proposés sur le nombre total de documents pertinents dans l'ensemble de données. Il évalue la capacité du modèle à récupérer des documents pertinents à partir de l'ensemble du jeu de données.

6.2. Génération de réponse

Généralement, plusieurs métriques clés sont utilisées pour la **génération** : elles évaluent la qualité de la sortie générée, en s'assurant de sa pertinence et de sa cohérence[17].

De plus, de nouveaux cadres tels que *Graph RAG* et *Video RAG* ont émergé. Ils intègrent des données structurées en graphes et du contenu vidéo afin d'améliorer la récupération de connaissances et la génération de réponses.

[17] Atamel.dev. Evaluating RAG pipelines. Atamel.dev post. January 9, 2025. https://atamel.dev/posts/2025/01-09_evaluating_rag_pipelines/

D'autres métriques évaluent la qualité de la réponse générée par rapport à la requête initiale ou les documents sources. Par exemple, *Evaluation of Retrieval-Augmented Generation: A Survey*[16] introduit les métriques suivantes :

- ***Relevance* ou Pertinence** (Réponse ↔ Requête) mesure l'adéquation du document généré avec l'intention et le contenu de la requête initiale. Elle garantit que la réponse est pertinente et répond aux exigences spécifiques de la requête.
- ***Faithfulness* ou Fidélité** (Réponse ↔ Documents pertinents) évalue si le document généré reflète avec précision les informations contenues dans les documents pertinents et mesure la cohérence entre le document généré et les documents sources.
- ***Correctness* ou Exactitude** (Réponse ↔ Exemple de réponse) est similaire à l'exactitude du composant de récupération, cette exactitude est mesurée par rapport à un exemple de réponse, qui sert de référence. Elle vérifie si la réponse est correcte en termes d'informations factuelles et appropriée au contexte de la requête.

6.3. Globales

Il y a ensuite des métriques qui permettraient d'évaluer la qualité globale d'une réponse, de la requête utilisateur initiale à la réponse générée par le RAG, qui sont nécessaires.

Les métriques suivantes sont fréquemment utilisées :

- ***ADA*** : mesure la similarité sémantique entre une réponse générée et une réponse de référence à l'aide d'*embeddings*. Un score élevé indique une forte similarité contextuelle.
- ***Fuzzysim* :** évalue la ressemblance textuelle en tolérant les variations orthographiques, les synonymes et les paraphrases.
- ***LLM-as-a-Judge*** : utilise un *LLM* comme « juge » pour évaluer la pertinence, l'exactitude et la cohérence d'une réponse.
- **BLEU** (Bi*Lingual Evaluation Understudy*) **:** mesure la correspondance de *n-grams* entre le texte généré et la ou les références (optimisée pour la traduction).

- **ROUGE** (*Recall-Oriented Understudy* for *Gisting Evaluation*) **:** évalue la couverture des informations clés (rappel de *n-grams*). Idéale pour les résumés.

D'autres métriques, plus spécifiques[16] peuvent être intéressantes à considérer :

- **Latence (*Latency*)** : mesure la rapidité avec laquelle le système trouve l'information et réagit, un facteur crucial pour l'expérience utilisateur.
- **Diversité (*Diversity checks*) :** vérifie si le système récupère une variété de documents pertinents et génère des réponses diversifiées.
- **Robustesse au bruit (*Noise Robustness*)** : évalue la capacité du système à traiter les informations non pertinentes sans affecter la qualité de la réponse.
- **Rejet négatif (*Negative Rejection*)** : évalue la capacité du système à s'abstenir de fournir une réponse lorsque les informations disponibles sont insuffisantes.
- **Robustesse contrefactuelle (*Counterfactual Robustness*)** : évalue la capacité du système à identifier et à ignorer les informations incorrectes, y compris en cas d'alerte concernant une désinformation potentielle.

Des exigences supplémentaires, telles que la lisibilité, la toxicité, la perplexité et d'autres, peuvent être nécessaires pour une application plus humaine.

6.4. Globales avec citation

En reprenant les jeux de données contenant :

- Question.
- Réponse attendue.
- Citations attendues.

Et les trois points d'évaluations :

1) ***Retrieval*** : documents retournés.
2) **Réponse** : qualité de la réponse de l'IA*G.*
3) **Citations** : complétude des réponses de l'IA*G.*

Et en reprenant la figure 9 « Schéma *RAG* citations retournées » on constate que :

- Le point « Citations retournées » est un prérequis pour la qualité de Réponse équivalente et Citations équivalentes.

- La perception de la qualité par l'utilisateur porte principalement sur le point *Réponse équivalente*. Une mauvaise gestion des citations par le *LLM* n'implique pas nécessairement que sa réponse soit erronée.
- Les points Citations *retournées* et C*itations équivalentes* sont liés et leur lien indique la capacité du *LLM* à gérer les citations.
- Si le point Citations *retournées* inclut de nombreux faux positifs, mais que les réponses *Réponse équivalente* et *Citations équivalentes* sont satisfaisantes, alors le *LLM* a une bonne capacité à ignorer les contenus non pertinents.

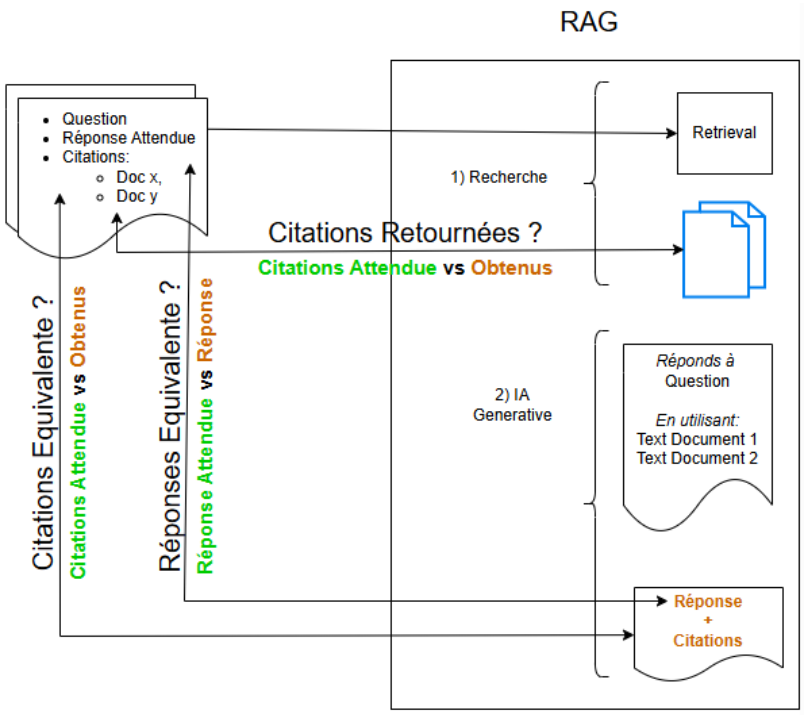

Figure 7 : Schéma *RAG* citations retournées

Formule proposée

Le groupe de travail avance sur une nouvelle métrique qui combinerait les différents aspects évoqués précédemment. Il nous parait en effet approprié de produire une unique formule qui donne un indicateur unique de qualité.

Cette formule pourrait par exemple se baser sur une combinaison de trois métriques, chacune avec une pondération associée. Par exemple :

```
ScoreRAGGlobal = (WRecherche × Score Recherche) +
                 (WGénération × Score Génération) +
                 (WCitation × Score Citation)
```

Où :

- **WRecherche, WGénération, WCitation** : sont les pondérations (***W**eight*) attribuées à chaque composant (par exemple, des valeurs entre 0 et 1, ou des pourcentages, dont la somme peut être 1 ou 100). Ces pondérations peuvent être ajustées en fonction des priorités du cas d'usage.
- **ScoreRecherche** : est le score d'évaluation de la Recherche Documentaire. Sans doute un score de Rappel ou un *F-Score*.
- **ScoreGénération** : est le score d'évaluation de la Génération de Réponse. Sans doute un score obtenu via un *LLM-as-a-Judge*.
- **ScoreCitation** : est le score d'évaluation des Citations. Sans doute un score de Précision.

6.5. Métriques des outils d'évaluation

Dans la section suivante, nous documentons une série d'outils d'évaluation des chaînes de *RAG*. Les métriques suivantes sont très largement utilisées par ces outils. À ce titre, elles constituent donc un **consensus du marché**.

Mean Reciprocal Rank (MRR)

La *Mean Reciprocal Rank (MRR)*, ou « moyenne du rang réciproque » en français, est une mesure statistique utilisée principalement pour évaluer la qualité des systèmes de recherche d'information ou de recommandation. Elle calcule la moyenne des inverses des rangs de la première réponse pertinente pour un ensemble de requêtes. Autrement dit, pour chaque requête, on prend le rang de la première bonne réponse, on en fait l'inverse (1/rang), puis on fait la moyenne de ces valeurs sur toutes les requêtes. Plus la *MRR* est élevée, meilleure est la performance du système, car cela signifie que les réponses pertinentes apparaissent plus tôt dans les résultats.

NDCG

La méthode *NDCG (Normalized Discounted Cumulative Gain)* utilisée par *Phoenix* pour l'évaluation des chaînes *RAG* sert à mesurer la qualité du classement des

documents récupérés par rapport à leur pertinence pour une requête donnée. Calcul du *NDCG@k* :

- On calcule le *DCG (Discounted Cumulative Gain*) qui prend en compte la pertinence des documents et leur position dans la liste (les documents les mieux classés comptent plus).
- On normalise ce score en le divisant par le score idéal (*IDCG*), c'est-à-dire le score que l'on obtiendrait si tous les documents les plus pertinents étaient en tête de liste.

Le NDCG varie entre 0 et 1 : plus il est proche de 1, meilleure est la qualité du classement.

Context relevance (pertinence du contexte)

C'est la mesure dans laquelle le contexte récupéré (c'est-à-dire les documents ou extraits fournis au modèle) est pertinent et utile pour répondre à la question posée. Un contexte est jugé pertinent s'il contient des informations qui aident directement à formuler une réponse correcte à la requête de l'utilisateur.

Groundedness (ancrage ou fidélité au contexte)

Cela évalue à quel point la réponse générée par le modèle est bien fondée sur le contexte fourni. Une réponse est dite « *grounded* » si elle s'appuie strictement sur les informations présentes dans le contexte récupéré, sans introduire d'éléments extérieurs ou d'hallucinations.

Retrieval quality

Désigne la qualité avec laquelle le système parvient à retrouver et fournir des documents ou extraits pertinents en réponse à une question donnée. Cette métrique évalue la pertinence, l'exactitude et l'utilité des documents récupérés pour aider à générer une réponse correcte. Une bonne qualité de récupération signifie que le système met à disposition des informations précises et adaptées, facilitant ainsi la production de réponses fiables et pertinentes.

Answer relevance (pertinence de la réponse)

Cette métrique mesure dans quelle mesure la réponse générée répond réellement à la question initiale de l'utilisateur. Une réponse pertinente traite directement la demande, en s'appuyant sur le contexte, et apporte une information utile et appropriée à la requête.

Hit Rate (taux de réussite)

C'est une métrique qui mesure la proportion de requêtes pour lesquelles au moins un document pertinent a été récupéré parmi les résultats proposés par le système. Autrement dit, pour chaque question, si le système parvient à retrouver au moins un document contenant l'information correcte ou utile, cela compte comme un "*hit*". Le *hit rate* est alors le pourcentage de ces succès sur l'ensemble des requêtes testées.

Par exemple, si sur 100 questions, le système retrouve au moins un document pertinent pour 85 d'entre elles, le *hit rate* sera de 85 %. Cette métrique permet d'évaluer l'efficacité du module de récupération à fournir des contextes pertinents pour la génération de réponses.

HUB FRANCE IA

7. Outils

7.Outils

7.1.Outils *Open Source*

Dans cette section, nous allons présenter les outils dont on dispose pour l'évaluation des *RAG*.

RAGAS

La librairie *RAGAS*[18] (*Retrieval-Augmented Generation Assessment*) propose une évaluation « simplifiée », utilisant la précision moyenne (*AP : Average Precision*) et des métriques personnalisées comme la « *Faithfulness* » (ou fidélité). Il évalue la pertinence du contenu généré par rapport aux contextes fournis et est adapté aux évaluations initiales ou lorsque les données de référence sont rares (donc lorsqu'on n'a pas forcément de vérité absolue définie par un humain).

Pour évaluer un *pipeline* de *RAG*, *RAGAS* attend les informations suivantes :

- **Question** : la requête de l'utilisateur (l'input de la « *pipeline* » de *RAG*).
- **Réponse** : la réponse générée par la « *pipeline* » (l'*output*).
- ***Contexts*** : les données contextuelles reçues grâce au corpus documentaire externe pour répondre à la question.
- ***Ground_truths*** : la « véritable » réponse à la question. Ce qui est intéressant est que cette information (qui est donc annotée par un humain) n'est pas obligatoire pour toutes les métriques mises à disposition par *RAGAS*.

RAGAS propose des métriques pour évaluer le composant « *Retrieval* » (*context_relevancy & context_recall*) mais aussi le composant « *Generation* » (*faithfulness & answer_relevancy*) séparément.

- ***Context_precision*** : mesure le ratio « *signal to noise* » du contexte récupéré. Cette métrique est calculée à partir de « question », des « *contexts* » et de la « réponse » OU des « *ground_truths* ».

[18] https://docs.ragas.io/en/latest/

- ***Context recall*** : mesure si toutes les informations pertinentes nécessaires à la réponse ont été récupérées. Cette métrique est calculée à partir de « *ground_truth* » et des « *contexts* ».
- ***Faithfulness*** : mesure la précision factuelle de la réponse générée. Le nombre d'affirmations correctes à partir des « *contexts* » est divisé par le nombre total d'affirmations dans la réponse. Cette métrique utilise « question » et « *contexts* ».
- ***Answer relevancy*** : mesure la pertinence de la réponse générée par rapport à la question. Cette métrique est calculée en utilisant « question » et « *answer* ».

Ces métriques concernent chacune des composants de la chaîne *RAG* individuellement, mais il existe également des métriques pour évaluer l'ensemble de la chaîne.

ARES

La méthode *ARES* (*Automated Retrieval-Augmented Generation Evaluation System*)[19] évalue les chaînes de *RAG* (ainsi que les applications *RAG*) en utilisant des techniques de génération de données synthétiques et en optimisant des modèles de classification afin de **minimiser au maximum le besoin d'annotations humaines**.

La méthodologie employée consiste à créer des requêtes synthétiques et des réponses à partir des documents, qui sont ensuite utilisées pour entraîner des « *lightweight language model* », qui agiront en tant que modèles « juges ».

Ces modèles « juges » attesteront de la qualité individuelle de chaque composant de la chaîne de *RAG*, en se basant sur des métriques telles que : *context relevance*, *answer faithfulness*, et *answer relevancy*.

[19] Jon Saad-Falcon, Omar Khattab, Christopher Potts, Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*. 2023. https://arxiv.org/pdf/2311.09476

ARES utilise le *Prediction-Powered-Inference*[20] (*PPI*) pour raffiner les évaluations, et fournir des informations statistiques additionnelles pour chaque résultat. Cette approche permet d'améliorer l'évaluation sur divers domaines, même avec peu de données annotées par un humain.

DeepEval

DeepEval [21] évalue les chaînes de *RAG* et les applications *RAG* en gérant séparément chaque composant de la chaîne (*Retrieval & Generation*). La méthodologie implique de créer des cas de tests avec des requêtes, les réponses attendues, et le contexte du *retriever*. Pour le *retriever, DeepEval* utilise des métriques telles que la précision contextuelle, le rappel et la pertinence afin d'évaluer la correspondance du contexte récupéré avec la requête.

Pour le *generator*, les métriques prises en compte sont : *answer relevancy* et *faithfulness*, afin de s'assurer de la validité des réponses générées. Cette approche permet de détecter rapidement les problèmes au niveau des composants de la chaîne, facilitant ainsi le *debugging* et l'optimisation.

Phoenix

La méthode *Phoenix*[22] évalue les chaînes de *RAG* en fournissant un *framework* pour évaluer à la fois le composant *retriever* et le composant *generator*. La méthodologie de *Phoenix* consiste à appliquer la chaîne de *RAG* pour récupérer des *logs* d'interactions et de *feedback*, qui sont ensuite analysés en calculant des métriques clés.

- Pour l'évaluation de la partie *retrieval, Phoenix* calcule les métriques *MRR*, *precision @K*, et *NDCG* pour mesurer à quel point le contexte est pertinent par rapport à la question initiale.

---

[20] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, Tijana Zrnic. Prediction-powered inference. *Science*, vol. 382, n°6671, p. 669-674. 2023. https://www.science.org/doi/10.1126/science.adi6000

[21] https://deepeval.com/

[22] https://arize.com/docs/phoenix/cookbook/evaluation/evaluate-rag

- Pour l'évaluation de la partie *generation*, les métriques utilisées sont *context relevance, groundedness,* et *answer relevancy.*

Phoenix intègre ces évaluations dans un système de *reporting* et dans des *dashboards* visuels.

7.2. Solutions commerciales

LangSmith

LangSmith[23] évalue les chaînes de *RAG* (ainsi que les applications *RAG*) en suivant une méthodologie structurée, incluant :

- Création de *datasets* de test (questions & réponses attendues).
- Lancement de la chaîne de *RAG* sur les *datasets* (génération de réponses).
- Mesure des performances en utilisant diverses métriques (*answer relevancy, accuracy, retrieval quality*).

En comparant les réponses générées et les réponses attendues, *LangSmith* permet d'obtenir des éléments permettant de savoir à quel point la chaîne de *RAG* est efficace pour récupérer et générer des informations pertinentes.

LangSmith propose un « *tier* » gratuit mais avec des fonctionnalités limitées.

RagMetrics

Cette startup[24] américaine, fondée par un entrepreneur français, propose une solution dédiée à l'évaluation des chaînes de *RAG*.

La mise en œuvre est assez simple : Il suffit d'exposer dans sa solution un *end point* qui, pour une question donnée, renvoie une réponse avec les éléments de contexte.

Ensuite, dans la solution *RagMetric*, il est possible de lancer des tests en *batch*, et d'obtenir un *score global* via une évaluation de type « *LLM-As-a-Judge* ».

[23] https://www.langchain.com/langsmith
[24] https://ragmetrics.ai

LlamaIndex

LlamaIndex [25] évalue les chaînes de *RAG* en permettant d'implémenter des modules d'évaluation pour les parties *Retrieval* et *Generation*. La méthodologie est la suivante :

- Créer des jeux de données synthétiques à partir d'un corpus de texte non structuré.
- Générer les *embeddings vectoriels*, et les stocker dans une base de données vectorielle.

Le *retriever* est ensuite évalué en utilisant des métriques de *ranking* comme *MRR* et *hit-rate*. Pour le *generator, LlamaIndex* utilise des métriques comme : *correctness*, *semantic similarity, faithfulness, context relevancy* et *answer relevancy*.

Ces métriques permettent de s'assurer que les réponses générées correspondent aux réponses attendues, respectent les *guidelines*, et sont fidèles au contexte récupéré. En intégrant des outils d'évaluation créés par la communauté comme *UpTrain*, *Tonic* et *DeepEval*, *LlamaIndex* propose un process d'évaluation robuste et flexible.

LlamaIndex propose un « *tier* » gratuit mais avec des fonctionnalités limitées.

TruLens

TruLens [26] évalue les chaînes de *RAG* en se concentrant sur trois métriques principales : *context relevance, groundedness*, et *answer relevancy*. La méthodologie est la suivante :

- Appliquer la chaîne de *RAG* sur un jeu de données.
- Sauvegarder les interactions et le *feedback* de la chaîne.

Les données récupérées sont ensuite analysées à l'aide des métriques décrites précédemment, *context relevance, groundedness* et *answer relevancy*.

---

[25] https://www.llamaindex.ai/
[26] https://www.trulens.org/

En mettant en œuvre ces évaluations, *TruLens* permet d'obtenir une compréhension nuancée des performances d'une application *RAG*. Il est possible de tester des *QA* et du *debugging* sur l'application *RAG*.

TruLens-Eval propose un « *tier* » gratuit mais avec des fonctionnalités limitées.

EvidentlyAI

EvidentlyAI[27] évalue les chaînes de *RAG* en fournissant un *framework* d'évaluation de la partie « *retrieval* » et de la partie « *generation* ».

La méthodologie consiste à créer des jeux de données synthétiques, comprenant des requêtes, des contextes récupérés lors de la partie « *retrieval* » et des réponses générées.

Pour l'évaluation du *retriever*, *EvidentlyAI* utilise les métriques « *context relevance* » et « *hit rate* » pour déterminer dans quelle mesure l'information récupérée correspond à la question initiale.

Pour l'évaluation du « *generator* », les métriques utilisées sont « *answer accuracy* », « *faithfulness* » et « *relevance* ».

EvidentlyAI intègre ces évaluations dans un système de *reporting* structuré, incluant notamment des *dashboards* visuels, ce qui simplifie le *monitoring* de la performance de la chaîne de *RAG*[28].

7.3. Matrice de comparaison

Afin de mieux appréhender les outils disponibles pour l'évaluation des chaînes *RAG*, cette section propose une analyse comparative des principales solutions du marché et des *frameworks open-source*. Ces outils couvrent différents cas d'usage – depuis l'évaluation ponctuelle d'un composant jusqu'au monitoring en production – et se distinguent par leurs méthodologies, leurs métriques d'évaluation, ainsi que leur niveau d'automatisation.

---

[27] https://www.evidentlyai.com/

[28] *EvidentlyAI* est une solution payante mais basée sur le *framework* open-source « *evidently* » https://github.com/evidentlyai/evidently

La matrice comparative présentée au tableau 2 « Matrice de comparaison des outils d'évaluation pour chaînes *RAG* » synthétise ces différences selon quatre dimensions principales : l'outil, son cas d'usage, son mode de fonctionnement (*open-source*, abonnement, type d'évaluation) et les métriques supportées (génération, récupération, fidélité, pertinence, etc.). Cette vue d'ensemble permet de choisir un outil adapté en fonction du contexte d'évaluation visé.

Outil	Use Case	Fonctionnement	Métriques
RAGAS	Évaluation de composants de la chaîne de *RAG* ou de la chaîne en entier, avec ou sans annotations humaines.	*Framework open-source*. Évaluation de la pertinence du contenu récupéré et généré par rapport à la question et aux contextes fournis.	• *Retrievalcontext_relevancy, context_recall* • *Generationfaithfulness, answer_relevancy*
LangSmith	Évaluation d'applications *RAG*	Tier gratuit avec fonctionnalités limitées, abonnement nécessaire pour des fonctionnalités avancées. Chaîne de traitement : création de *datasets* de test avec questions et réponses attendues, application du *RAG* sur ces *datasets*, et évaluation des performances.	*Answer relevancy, answer accuracy, retrieval quality*

***ARES* (Automated *RAG* Evaluation System)**	Évaluations de chaînes de *RAG* dynamiques et continues	*Framework* open-source. Utilisation combinée de génération de données synthétiques et de modèles de classifications fine-tunés pour évaluer la pertinence de contexte, la fidélité et la pertinence de la réponse, le tout en minimisant le besoin de vérité absolue (annotations humaines).	*Mean Reciprocal Rank, Normalized Discounted Cumulative Gain*
DeepEval	Évaluation de *LLMs*, tests similaires à des tests unitaires	*Framework* open-source. Évaluation des composants du *RAG* séparément.	*Retrieval*: *Precision, Recall, Pertinence.* *Generation*: *Pertinence, Faithfulness.* *RAG Triad.* Métrique *GEval*, permet d'adapter l'évaluation avec des critères personnalisés !
LlamaIndex	Construction & évaluation d'applications *LLMs.*	*Framework* incluant des outils pour générer des *datasets* de test synthétiques (ou non), et faire de l'évaluation. Génère des questions automatiquement	*Context precision, recall, faithfulness, response relevancy.*

		à partir des données. Tier gratuit, mais fonctionnalités additionnelles nécessitent un abonnement.	
TruLens	Évaluer la qualité et l'efficacité en temps réel d'applications basées sur des *LLMs.*	Intègre plusieurs *frameworks* pour calculer des métriques et réaliser du *QA / debugging.* Tier gratuit mais fonctionnalités additionnelles nécessitent un abonnement.	*RAG Triad*: *context relevancy, groundedness,* et *answer relevancy*.
EvidentlyAI	Évaluation, test, monitoring de systèmes IA.	Permet de créer un système de *reporting* structuré, et d'implémenter des *dashboards* de monitoring de la performance de la chaîne. Tier gratuit mais avec fonctionnalités limitées, solution basée sur le *framework open-source « evidently* ».	Pertinence de contexte (même au niveau *chunk*), *ranking* (*hit rate*), qualité de génération avec ou sans vérité absolue, utilisation de *LLMs « juges ».*
Phoenix	Évaluation flexible d'applications *LLM.*	Métriques pour *QA et debugging*. Système de *reporting*	*Context precision, recall, faithfulness, response relevancy*.

		structuré et *dashboards* de monitoring. Tier gratuit mais fonctionnalités additionnelles nécessitent un abonnement.	

Tableau 2 : Matrice de comparaison des outils d'évaluation pour chaînes *RAG*

Plusieurs métriques sont pertinentes, notamment celles utilisant un *juge LLM,* avec une instruction initiale et des exemples limités (*few shot examples*) pour chaque métrique. C'est le cas, par exemple, de la métrique « *Context Precision* » :

- `Instruction: Given question, answer and context verify if the context was useful in arriving at the given answer. Give verdict as "1" if useful and "0" if not with json output.`
- `Examples (questions + contexte + réponse) : examples=[(QAC(`
    - `question='What can you tell me about Albert Einstein?',`
    - `context="Albert Einstein (14 March 1879 – 18 April 1955) was a German-born theoretical physicist, widely held to be.....",`
    - `answer='Albert Einstein, born on 14 March 1879, was a German-born theoretical physicist, widely held to be.....')`
- `Vérifications: Verification(reason="The provided context was indeed useful in arriving at the given answer. The context includes key information about Albert Einstein's life and contributions, which are reflected in the answer.", verdict=1))`

Il est important de noter que ces exemples sont initialement tous en anglais. Il est donc nécessaire d'adapter certaines métriques lorsqu'on souhaite les appliquer sur des jeux de données d'une autre langue. Cette adaptation consiste à traduire chaque exemple et chaque vérification vers la langue cible, tandis que l'instruction initiale reste en anglais. Sur un jeu de données en français, les résultats sont effectivement bien meilleurs après avoir adapté chaque métrique pour traduire les exemples automatiquement.

HUB FRANCE IA

8. Résultats et analyse

8. Résultats et analyse

Le groupe de travail débute l'analyse des premiers résultats d'évaluations au moment de cette publication. Vous ne trouverez donc, pour l'instant, qu'une seule analyse.

Si vous souhaitez partager vos résultats, que nous intégrerons à nos analyses, merci de nous contacter en envoyant un email à alberto.tepox@hub-franceia.fr

8.1. Évaluation de la similarité attendu *vs* obtenu

L'évaluation de chaîne de *RAG* a été réalisée sur le *Single-Topic-RAG Evaluation Dataset* à l'aide de diverses métriques d'évaluation, parmi lesquelles les plus significatives furent *Ada, FuzzySim, LLM as Judge* (DeepSeek-V3-0324: classique et : *reasoner*) ainsi qu'une métrique définie via une formule : *max_jarowinkler_ratcliffobershelp_ada* qui prend la similarité maximum de deux algorithmes de *NLP, RatcliffOberhelp* et *JaroWinkler,* ainsi qu'une similarité vectorielle sur le modèle *Ada.* Le principe consistait à comparer la réponse d'un *LLM* à la réponse attendue en utilisant différentes métriques. Pour chaque réponse générée par le *LLM*, un évaluateur humain a attribué une note de 0 à 10 reflétant la similarité avec la réponse de référence. Les valeurs obtenues par chaque métrique ont ensuite été comparées aux notes humaines afin d'identifier la métrique présentant la meilleure corrélation avec l'évaluation humaine.

Voici les résultats les plus intéressants de notre étude :

- Classement par corrélation avec l'évaluation humaine
    - *Reasoner LLM* (DeepSeek-R1-0528): 0.869 (meilleure performance globale).
    - *LLM* Classique (DeepSeek-V3-0324) : 0.788.
    - *FuzzySim* : 0.719 (meilleure métrique algorithmique).
    - max_jarowinkler_ratcliffobershelp_ada : 0.506.
    - *Ada* : 0.504.
- Top 3 des métriques algorithmiques
    - FuzzySim : 0.719.

- max_jarowinkler_ratcliffobershelp_ada : 0.506.
- *Ada* : 0.504.

8.2. Analyse des performances

Le *Reasoner LLM* démontre un score de 10.3% supérieur au *LLM* classique dans sa capacité à reproduire le jugement humain. Cette amélioration s'explique par le processus de raisonnement explicite intégré dans *Deepseek-Reasoner*, qui permet une évaluation plus nuancée et contextualisée des réponses. Les performances relatives des différentes approches sont synthétisées dans la matrice de corrélation présentée au Tableau 3 « Matrice de corrélation entre **évaluation** humaine, juge *LLM* et métriques de similarité ». Celle-ci met en évidence les écarts de corrélation entre l'évaluation humaine, les juges *LLM* et les principales métriques de similarité, permettant ainsi de visualiser clairement la supériorité des modèles à raisonnement explicite.

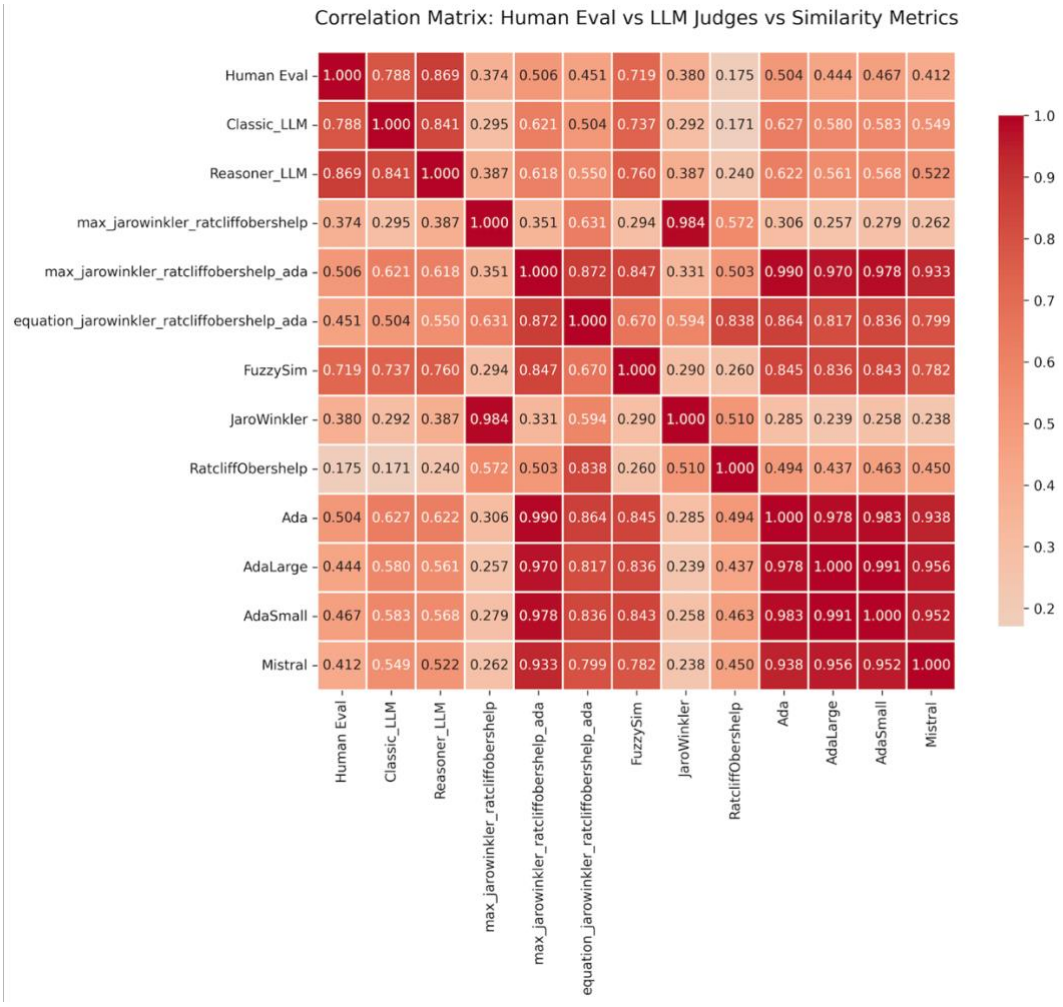

Tableau 3 : Matrice de corrélation entre évaluation humaine, juge *LLM* et métriques de similarité

Comme représenté dans le tableau 3, en observant les valeurs de corrélations entre chaque métrique et l'évaluation humaine (*Human Eval*) les deux approches *LLM* surpassent significativement les métriques de similarité algorithmiques traditionnelles (qui, à l'exception notable de *FuzzySim*, ont toutes des scores inférieurs à 50% en comparaison avec l'évaluation humaine). D'après le tableau 3, l'évaluation par le *Reasoner LLM* présente une similarité avec l'évaluation humaine 20.9% supérieure à la meilleure métrique algorithmique (*FuzzySim*), ce qui confirme la valeur ajoutée des modèles de langage avancés pour l'évaluation automatique de systèmes *RAG*.

8.3. Implications pratiques

Ces résultats suggèrent que l'utilisation d'un *LLM* avec capacité de raisonnement explicite comme juge automatique constitue une approche plus fiable pour l'évaluation de chaînes *RAG*, en particulier dans des contextes où l'alignement avec le jugement humain est crucial pour la validation de la qualité des réponses générées.

HUB FRANCE IA

CONCLUSION

Conclusion

Nous concluons ce guide pratique avec une certaine frustration, n'ayant pas encore pu collecter suffisamment de résultats pour formuler des observations plus complètes et des recommandations d'optimisation pour les chaînes de *RAG*. Nous tenons à préciser que nous envisageons de mettre à jour ce document, y compris cette conclusion, dans une deuxième édition, en y intégrant une section « Résultats » qui propose des éléments plus détaillés.

Il nous semble que la **problématique de l'évaluation** demeure fondamentale pour le déploiement de l'IAG. En effet, la confiance naturelle que les modèles peuvent inspirer en annonçant des contre-vérités doit absolument être contrebalancée par des métriques fiables permettant à chacun de faire la part des choses et de mettre en place les garde-fous nécessaires. À cet égard, l'évaluation revêt une importance capitale tant pour l'acculturation des utilisateurs que pour l'optimisation des chaînes.

La **citation des sources** est une bonne pratique du journalisme ou de la recherche, permettant de limiter considérablement les *fake news* (humaines) ou les hallucinations (IAG). À ce titre, il nous semble important, comme évoqué dans nos propositions, de poursuivre les travaux sur l'évaluation des réponses, en s'appuyant sur les citations afin d'identifier une métrique robuste en la matière et de repérer les citations erronées également.

Ce document a mis en évidence la difficulté d'une évaluation fiable des réponses de l'IAG. Nous sommes parvenus à un stade de développement tel qu'une telle évaluation n'est pas plus aisée que l'évaluation d'une tâche de *NLP* classique en français (résumé, analyse de texte, etc.) qui présente elle-même des difficultés de mise en œuvre. En effet, tant les élèves que le corps enseignant soulignent une problématique de subjectivité : une même copie de français au collège ne reçoit pas la même note de la part de tous les enseignants. De plus, la relecture et l'évaluation représentent un investissement en temps considérable … Imaginer que l'évaluation de l'IAG serait plus simple que l'évaluation de la copie d'un

étudiant, et pourrait être entièrement automatisée, sans limites, est un leurre à notre sens.

Et ainsi, malgré ses meilleurs taux de pertinence pour l'évaluation, l'emploi de la méthode « *LLM As Judge* », bien que très intéressante, ne peut être pleinement satisfaisante. Un modèle ne peut être juge et partie.

HUB FRANCE IA

RÉFÉRENCES

Références

De nombreux documents de recherche et articles de blog traitent de ce sujet. Les documents suivants ont été examinés attentivement et utilisés dans notre état de l'art, constituant la première phase de nos travaux de recherche. En plus des références citées au cours du texte, on pourra consulter les références suivantes pour approfondir :

- Assaf Pinhasi. Evaluating RAG for large scale codebases. Blog Qodo.ai. 2025. https://www.qodo.ai/blog/evaluating-rag-for-large-scale-codebases/
- Atamel.dev. Evaluating RAG pipelines. Atamel.dev post. January 9, 2025. https://atamel.dev/posts/2025/01-09_evaluating_rag_pipelines/
- Nathan Atox, Mason Clark. LLMs. Evaluating Large Language Models through the Lens of Linguistic Proficiency and World Knowledge: A Comparative Study. *Authorea Preprints*. August 2024. https://www.authorea.com/doi/full/10.22541/au.172479372.22580887
- Ingeol Baek, Hwan Chang, Byeongjeong Kim, Lee, imin Lee, Hwanhee Lee. Probing-RAG: Self-Probing to Guide Language Models in Selective Document Retrieval. *arXiv preprint arXiv:2410.13339*, 2024. https://arxiv.org/pdf/2410.13339
- Nicolas Cavallo. Évaluation RAG : Bonnes pratiques pour assurer la mise en production. Blog OCTO, 21 mars 2025. https://blog.octo.com/evaluation-rag-bonnes-pratiques-pour-assurer-la-mise-en-production
- Eden AI. Le guide 2025 de la génération augmentée par récupération (RAG). Tutoriel. 2025. https://www.edenai.co/fr/post/the-2025-guide-to-retrieval-augmented-generation-rag
- Eval4RAG. Workshop on Evaluation of RAG Systems, ECIR 2025, 10 avril 2025. https://eval4rag.github.io/ et https://login.easychair.org/cfp/Eval4RAG
- Kurt Muehmel. The LLM Mesh. O'Reilly Media. November 2025. https://www.oreilly.com/library/view/the-llm-mesh/9781098176631/
- Reginald Martyr. Mastering RAG Evaluation: Best Practices & Tools for 2025. Orq.ai. November 29, 2024. https://orq.ai/blog/rag-evaluation

- Shangeetha Sivasothy, Scott Barnett, Stefanus Kurniawan, Zafaryab Rasool, Rajesh Vasa. RAGProbe: An Automated Approach for Evaluating RAG Applications. *arXiv preprint arXiv:2409.19019.* 2024. https://arxiv.org/abs/2409.19019
- Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi, Daniel Ho, James Zou, J. (2024). How well do LLMs cite relevant medical references? An evaluation framework and analyses. *arXiv preprint arXiv: 2402.02008*, 2024. https://arxiv.org/abs/2402.02008
- Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, Zhaofeng Liu. Evaluation of Retrieval-Augmented Generation: A Survey. In *CCF Conference on Big Data*. Singapore: Springer Nature Singapore, 2024. p. 102-120. https://arxiv.org/abs/2405.07437

HUB FRANCE IA

GLOSSAIRE

Evaluation des chaines de RAG

Septembre 2025

Glossaire

Acquired Podcast Transcripts	Ensemble de transcriptions de podcasts, questions-réponses et évaluations humaines, utilisé pour tester la performance des chaînes de *RAG* sur des contenus audio transcrits.
Answer Relevancy	Pertinence de la réponse générée par rapport à la question.
Approches hybrides	Combinaison de plusieurs stratégies pour optimiser la précision et le rappel.
BLEU	Mesure la correspondance de *n-grams* entre un texte généré et la référence (surtout pour la traduction).
Chaîne de *RAG* (*Retrieval-Augmented Generation*)	Ensemble de traitements combinant la recherche documentaire (*retrieval*) et la génération de texte par IA (générative). Elle permet de répondre à une requête utilisateur en injectant un contexte documentaire dans le *prompt* envoyé au *LLM*, afin d'exploiter des bases documentaires spécifiques plutôt que le savoir global du modèle.
Chunking	Découpage d'un document en unités plus petites (pages, paragraphes, passages) pour faciliter l'indexation et la recherche.
Citation	Référence explicite à la source documentaire utilisée pour générer une réponse, essentielle pour l'explicabilité et la vérification des réponses. Une citation est une référence à un Document Pertinent utilisé par le *LLM* pour construire sa réponse.
Context Precision (Précision du Contexte)	Ratio signal/bruit du contexte récupéré.

Context Recall (Rappel du Contexte)	Capacité à récupérer toutes les informations pertinentes nécessaires à la réponse.
Counterfactual Robustness	Capacité à ignorer ou corriger des informations incorrectes.
Diversity	Diversité des documents ou réponses générés.
Documents candidats	Ensemble de documents susceptibles d'être pertinents pour une requête.
Documents pertinents (*Relevant Documents*)	Documents effectivement utiles pour répondre à la question.
Elastic Search	Moteur d'indexation par *tokens*, recherche lexicale rapide (*BM25*, *TF-IDF*). *Elastic* est un outil robuste et flexible qui permet de définir une chaîne de prétraitement pour l'indexation (découpage et normalisation des *tokens*) et de nombreuses stratégies de recherches. Ce moteur de recherche est particulièrement adapté à la recherche de mots clés précis, tels que des codes d'erreur ou des références de produits.
Evaluable Output	Sortie générée par le système, pouvant être évaluée selon des critères définis.
Faithfulness	Fidélité factuelle de la réponse par rapport au contexte fourni.
Fuzzysim	Évalue la ressemblance textuelle en tolérant des variations orthographiques, synonymes et paraphrases.
Génération (*Generation*)	Phase où le *LLM* produit une réponse à partir de la question et du contexte documentaire extrait lors du *retrieval.*

Graph Database	Indexation basée sur les relations entre entités, recherche par parcours de graphe.
Hallucination	Réponse d'un *LLM* qui n'est pas avérée, souvent fausse et sans source de référence.
Indexation	Processus de préparation et d'organisation des documents pour permettre leur recherche efficace.
Jeu de test (*Test Set*)	Ensemble structuré de documents, questions, réponses attendues et citations attendues, utilisé pour évaluer la performance d'une chaîne de *RAG*.
Jeux de données de référence *(Ground Truth)*	Réponse ou information de référence considérée comme correcte, servant de base à l'évaluation des réponses générées. Dans le cadre de ce document, Les jeux de tests forment la *Ground Truth*.
Latency	Temps de réponse du système. Sur les *LLMs*, les principaux indicateurs de *Latency* sont les : • TTFT: Time to First Token • TTNT: Time to Next Token
LLM (*Large Language Model*)	Modèle de langage de grande taille, entraîné sur de vastes corpus de textes, capable de générer, résumer ou analyser du texte en langage naturel.
LLM as-a-Judge	Utilisation d'un *LLM* pour évaluer la pertinence, l'exactitude et la cohérence d'une réponse.
Métriques d'évaluation	Indicateurs quantitatifs permettant de mesurer la performance d'une chaîne de *RAG*.
Moteur de Recherche	Outils capables de remonter des documents (pertinents dans l'idéal) sur la base d'une requête textuelle. Dans le cadre des chaînes de *RAG*, la partie *Retrieval* est assuré par un moteur de recherche, souvent de type *Vector Store.*

Negative Rejection	Capacité à ne pas répondre en l'absence d'information suffisante. Cet indicateur est approprié pour évaluer la capacité du *RAG* à contrer les hallucinations
Noise Robustness	Résistance du système à l'information non pertinente.
Pipeline	Chaîne de traitements successifs appliqués aux données dans le cadre d'une application *RAG*.
Pré-traitement	Ensemble des opérations appliquées aux documents avant indexation : extraction, conversion, correction, normalisation, découpage (*chunking*).
Précision@k	Proportion de documents pertinents parmi les k premiers résultats.
Rappel@k	Capacité à retrouver tous les documents pertinents dans les k premiers résultats.
Retrieval	Phase de recherche documentaire dans la chaîne de *RAG*, consistant à identifier les documents les plus pertinents pour une requête donnée.
ROUGE	Métrique d'évaluation de la couverture des informations clés (rappel de n-*grams*, utile pour les résumés).
Single-Topic RAG Evaluation Dataset	Jeux de test structurés avec textes, questions, réponses et citations, permettant d'évaluer la capacité d'un système RAG à répondre à des questions sur un sujet donné.
Vector Store	Moteur d'indexation par vectorisation des contenus (*embeddings*), permettant la recherche par similarité sémantique (*cosinus*). Il existe de nombreux *Vector Store*, commerciaux ou *open*-source.

HUB FRANCE IA

REMERCIEMENTS

Remerciements

La définition du sujet et les premiers travaux ont été menés par Cédric Lopez d'Emvista et Amédée Potier de Konverso. Ce sujet a donné lieu à la création du groupe de travail « Évaluation des Chaînes de *RAG* » qui a travaillé en commun avec le groupe de travail « Voix et langage ».

Le Hub France IA remercie l'ensemble des participants aux groupes de travail, et tout particulièrement les contributeurs de ce livrable.

Les pilotes et le référent du GT

- Alban Petit, Konverso.
- Alberto Tépox, Hub France IA.
- Amédée Potier, Konverso.

Les contributeurs

- Alban Petit, Konverso.
- Amédée Potier, Konverso.
- Julien Raige-Verger, Médiamétrie.
- Nicolas Pierrot, Konverso.
- Thibault Chazal, Digital Product Studio.

Relecture

- Alberto Tépox, Hub France IA.
- Cédric Lopez, Emvista.
- Didier Schwab, Université Grenoble Alpes.
- Karel Bourgois, Voxist - Co-pilote GT Voix et langage.

Validation

- Caroline Chopinaud, Hub France IA.
- Françoise Soulié-Fogelman, Hub France IA.
- Pierre Monget, Hub France IA.

La touche finale

- Mélanie Arnould, Hub France IA.

GUIDE PRATIQUE

ÉVALUATION DES CHAÎNES DE RAG

Septembre 2025

HUB FRANCE IA