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

Vers une évaluation rigoureuse des systèmes RAG : le défi de la due diligence

Grégoire Martinon[1] Alexandra Lorenzo de Brionne[2] Jérôme Bohard[1]
Antoine Lojou[1] Damien Hervault[1] Nicolas Brunel[1, 3]
(1) Capgemini Invent France, 147 Quai du Président Roosevelt, 92130, Issy-les-Moulinaeux, France
(2) DiaDeep, 35 Rue Louis Guérin, 69100, Villeurbanne, France
(3) LaMME, ENSIIE, Université Paris-Saclay, 3 rue Joliot Curie, Bâtiment Breguet, 91190 Gif-sur-Yvette, France
gregoire.martinon@capgemini.com, a.lorenzo@diadeep.com
jerome.bohard@capgemini.com, antoine.lojou@capgemini.com,
damien.hervault@capgemini.com, nicolas.brunel@capgemini.com

RÉSUMÉ

L'IA générative se déploie dans des secteurs à haut risque comme la santé et la finance. L'architecture RAG (Retrieval Augmented Generation), qui combine modèles de langage (LLM) et moteurs de recherche, se distingue par sa capacité à générer des réponses à partir de corpus documentaires. Cependant, la fiabilité de ces systèmes en contextes critiques demeure préoccupante, notamment avec des hallucinations persistantes. Cette étude évalue un système RAG déployé chez un fonds d'investissement pour assister les due diligence. Nous proposons un protocole d'évaluation robuste combinant annotations humaines et LLM-Juge pour qualifier les défaillances du système, comme les hallucinations, les hors-sujets, les citations défaillantes ou les abstentions. Inspirés par la méthode Prediction Powered Inference (PPI), nous obtenons des mesures de performance robustes avec garanties statistiques. Nous fournissons le jeu de données complet. Nos contributions visent à améliorer la fiabilité et la scalabilité des protocoles d'évaluations de systèmes RAG en contexte industriel.

ABSTRACT

Towards a rigorous evaluation of RAG systems : the challenge of due diligence

The rise of generative AI, has driven significant advancements in high-risk sectors like healthcare and finance. The Retrieval-Augmented Generation (RAG) architecture, combining language models (LLMs) with search engines, is particularly notable for its ability to generate responses from document corpora. Despite its potential, the reliability of RAG systems in critical contexts remains a concern, with issues such as hallucinations persisting. This study evaluates a RAG system used in due diligence for an investment fund. We propose a robust evaluation protocol combining human annotations and LLM-Judge annotations to identify system failures, like hallucinations, off-topic, failed citations, and abstentions. Inspired by the Prediction Powered Inference (PPI) method, we achieve precise performance measurements with statistical guarantees. We provide a comprehensive dataset for further analysis. Our contributions aim to enhance the reliability and scalability of RAG systems evaluation protocols in industrial applications.

MOTS-CLÉS : LLM, RAG, hallucinations, annotations, LLM-Juge, due diligence.

KEYWORDS: LLM, RAG, hallucinations, annotations, LLM-as-Judge, due diligence.

Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM), pages 291–308.

ARTICLE : **Accepté à** EvalLLM2025 : Atelier sur l'évaluation des modèles génératifs (LLM) et challenge.

1 Introduction

Depuis l'apparition de ChatGPT, l'usage de l'IA générative s'est rapidement étendu à des secteurs sensibles comme la santé, la finance ou la défense. Parmi les applications notables, les systèmes RAG (Retrieval-Augmented Generation), combinant LLM (Large Language Model) et moteur de recherche, se distinguent par leur capacité à générer des réponses fondées sur un corpus documentaire, avec citations à l'appui.

Ces systèmes sont déjà largement adoptés dans l'industrie, notamment pour interroger des documentations d'entreprises complexes ou analyser des archives dans le cadre de fusions-acquisitions. Mais dans des contextes critiques, leur fiabilité est un enjeu crucial (Weidinger *et al.*, 2025; Zhou *et al.*, 2024). Malgré leurs promesses, les RAGs ne garantissent pas l'absence d'hallucinations (Magesh *et al.*, 2024), et leurs performances varient fortement selon le domaine métier. Cela impose des protocoles d'évaluation adaptés, en cohérence avec des exigences réglementaires croissantes, comme celles de l'AI Act.

Deux approches d'évaluation dominent : l'annotation humaine, précise mais coûteuse, et l'évaluation par LLM-Juge, scalable mais parfois biaisée. Des méthodes hybrides comme PPI (Prediction Powered Inference) (Angelopoulos *et al.*, 2023a,b; Boyeau *et al.*, 2024) ou ASI (Active Statistical Inference) (Zrnic & Candès, 2024; Gligorić *et al.*, 2024) proposent de les combiner pour une évaluation rigoureuse et économiquement viable. Dans cet article, nous exploitons ces méthodes pour évaluer un système RAG utilisé lors de due diligence. Nous développons un protocole approfondi, inspiré de PPI, pour quantifier les points de défaillance d'un système industriel [1].

Nos contributions principales sont les suivantes :

- Un protocole d'évaluation détaillé, par domaine métier, des performances d'un RAG industrialisé (hallucinations, hors-sujets, langue, citations, abstentions).
- Une prise en compte explicite de la température non nulle via des répétitions multiples.
- Un protocole hybride combinant annotations humaines et LLM-Juge, inspiré de PPI.
- Un jeu de données complet en français (questions, réponses, sources, annotations humaines et LLM-Juge), anonymisé et accessible publiquement : `https://github.com/gmartinonQM/eval-rag`

Dans cet article, nous retenons la définition d'une hallucination comme étant un élément non-déductible d'une base de connaissances vérifiable (Maleki *et al.*, 2024).

1. De récents travaux proposent de travailler à la maille du fait atomique généré dans le texte (Min *et al.*, 2023; Scirè *et al.*, 2024). Cependant, nous avons constaté que cette maille était extrêmement laborieuse à extraire pour les annotateurs humains et difficilement automatisable pour les LLMs, avec des taux d'oublis de l'ordre de 50%. Nous faisons donc le choix pragmatique de travailler à l'échelle de la phrase pour quantifier et localiser les hallucinations.

2 Travaux connexes

L'évaluation des LLMs est cruciale, notamment dans les systèmes à haut risque (Weidinger *et al.*, 2025; Zhou *et al.*, 2024). Les benchmarks comme HELM (Liang *et al.*, 2023) s'appuient sur des jeux de données labellisés, incluant des ensembles spécialisés pour détecter les hallucinations. ANAH (Ji *et al.*, 2024) et HaluEval (Li *et al.*, 2023) les caractérisent à la maille phrase dans des contextes généraux, là où TofuEval (Tang *et al.*, 2024) se focalise sur les résumés de dialogue. RAGTruth (Niu *et al.*, 2024) descend à la maille fait et se réfère à des passages du jeu de données MS Marco (Nguyen *et al.*, 2016), tandis que LLM-OASIS (Scirè *et al.*, 2024) s'intéresse aux faits en rapport avec Wikipédia. Notre jeu de données s'inscrit dans la lignée de MEMERAG (Blandón *et al.*, 2025), qui évalue les hallucinations et les hors-sujets à l'échelle de la phrase dans un contexte RAG, mais avec un ancrage industriel plus fort dans la due diligence.

Les jeux de données standards servent souvent à entraîner des détecteurs d'hallucinations, mais restent limités pour évaluer un système en contexte réel. Les méthodes basées sur un LLM-Juge (Zheng *et al.*, 2023) ou un comité de LLM-Juges (Chern *et al.*, 2024; Jung *et al.*, 2025), comme RAGAS (Es *et al.*, 2024) ou G-Eval (Liu *et al.*, 2023) permettent une évaluation automatique et scalable. Ces méthodes peuvent être enrichies d'outils tels que Google Search ou des interpréteurs Python, comme dans les approches SAFE (Wei *et al.*, 2024) et FactTool (Chern *et al.*, 2023). Certaines méthodes exploitent la stochasticité des LLMs pour estimer des scores d'hallucinations, c'est le cas de ChainPoll (Friel & Sanyal, 2023) ou des approches par entropie sémantique (Farquhar *et al.*, 2024). D'autres méthodes vont plus loin, et calibrent ces scores pour refléter la probabilité d'occurrence d'une hallucination (Valentin *et al.*, 2024). Cependant, ces méthodes sont sensibles au prompt, au modèle utilisé, et à la qualité des données disponibles, comme les graphes de connaissance (Mountantonakis & Tzitzikas, 2023). Les résultats obtenus par les LLM-Juges sont très variables, avec des performances oscillant entre 5% et 50% (Hong *et al.*, 2024).

À l'opposé du spectre, les annotations humaines offrent une grande précision mais sont coûteuses et difficilement scalables (Min *et al.*, 2023). Des méthodes hybrides comme PPI (Angelopoulos *et al.*, 2023a,b) et ASI (Zrnic & Candès, 2024) combinent les deux approches. Elles incarnent une théorie des sondages augmentée par LLM. PPI repose sur un échantillon de contrôle aléatoire annoté par des humains et LLM-Juge. Les annotations humaines sont utilisées pour corriger les biais du LLM-Juge, et appliquer cette correction à grande échelle. ASI affine encore cette logique via un échantillonnage adaptatif basé sur l'incertitude du LLM-Juge. PPI a été appliqué à des systèmes de compétition entre LLMs, tels que ChatbotArena (Boyeau *et al.*, 2024), ou à des systèmes RAGs (Saad-Falcon *et al.*, 2024). De son côté, ASI a été utilisé pour des systèmes de classification multi-classe basés sur des LLMs (Gligorić *et al.*, 2024). Nos travaux s'inscrivent dans la lignée de ARES (Saad-Falcon *et al.*, 2024), en exploitant l'approche PPI dans le contexte industriel de la due diligence.

3 Système évalué

3.1 Contexte industriel

Dans cet article, nous nous concentrons sur l'évaluation d'Alban, un assistant virtuel développé pour accompagner un fonds d'investissement international gérant plusieurs milliards d'euros d'actifs. Ce type d'acteur mène régulièrement des opérations de due diligence, c'est-à-dire un processus d'analyse

approfondie préalable à une décision d'investissement, de fusion ou d'acquisition. L'objectif est d'évaluer de manière rigoureuse la situation financière, juridique, opérationnelle et stratégique d'une entreprise cible, afin d'identifier les risques potentiels, de confirmer la valeur réelle de l'actif, et de négocier au mieux les termes de la transaction. Cette démarche repose sur l'étude de très nombreux documents internes (rapports d'activité, états financiers, contrats, procès-verbaux, audits, etc.) produits par l'entreprise analysée et transmis dans une *data room* dédiée.

Pour un fonds de grande envergure, ce processus représente un effort considérable : plusieurs millions d'euros sont investis à chaque due diligence dans l'analyse documentaire, avec un volume dépassant les 100 000 documents à examiner en quelques semaines seulement. Alban a été conçu pour rationaliser cette phase, en offrant aux équipes de transaction un agent conversationnel capable de répondre rapidement et précisément à leurs questions, à partir du corpus documentaire fourni.

La solution repose sur GPT-4, utilisé avec une température non nulle, et est aujourd'hui en production avec plus de 50 utilisateurs actifs.

3.2 Système RAG

Alban est un système RAG (Retrieval-Augmented Generation) de type "Advanced" (Gao *et al.*, 2023). Il commence par ingérer le patrimoine documentaire de l'entreprise, tous formats confondus (PDF, PowerPoint, Word, Excel), en le découpant en segments appelés sources. Chaque source est enrichie de métadonnées (titre, page) et vectorisée à l'aide d'un moteur d'embeddings, puis stockée dans une base vectorielle.

Lors d'une requête, le prompt utilisateur est reformulé, vectorisé et comparé à la base pour extraire les 15 sources les plus proches. Ces sources, combinées à la requête, sont fournies au LLM via un système de prompt en anglais pour générer une réponse pertinente.

La figure 1 illustre un exemple d'échange avec Alban.

Question
Comment DataCorp évalue-t-elle la satisfaction et l'expérience des salariés ?
Réponse
DataCorp utilise des enquêtes de satisfaction pour obtenir des commentaires sur l'engagement, le moral et la satisfaction des employés au travail. Ces enquêtes sont menées dans le cadre des dispositifs convenus avec ComeToMyCorp, une entreprise spécialisée dans l'évaluation de la satisfaction des employés [^5f7cce^]. Il est à noter que la direction de DataCorp a également mis en place des actions pour répondre aux commentaires laissés sur la plateforme JobReview, afin de gérer l'e-réputation de l'entreprise [^4ca822^][^63fadb^].

FIGURE 1 – Exemple de réponse générée par le système Alban. Les identifiants entre crochets correspondents aux identifiants des sources sur lesquelles le système base sa réponse.

4 Points de défaillance du système

En inspectant manuellement les réponses du système Alban, plusieurs points de défaillance ont été recensés :

— **Réponses stochastiques** : Le système peut donner une réponse différente si la même question lui est posée plusieurs fois.
— **Langue erronée** : Le système peut répondre en anglais, même si l'utilisateur écrit en français.
— **Réponses inattendues** : Le système peut répondre à des questions pour lesquelles aucune donnée n'est disponible.
— **Citations défaillantes** : Le modèle peut mal restituer les identifiants des sources, par exemple en citant `4ca823` au lieu de `4ca822`.
— **Hors-sujets** : Le système peut fournir des informations sans rapport avec la question posée.
— **Langue de bois** : Le système peut inclure des éléments non-assertifs, par exemple de pure politesse, ou des conjectures non vérifiables.
— **Hallucinations** : Le système peut inclure des éléments qui ne sont pas déductibles des sources citées.
— **Réponses partielles** : Le système peut fournir une réponse partielle, même lorsqu'il dispose de tous les éléments nécessaires dans les sources.

Dans le contexte de la due diligence, ces points de défaillance ont des impacts qui peuvent aller de la simple perte de temps (langue erronée, hors-sujets, langue de bois), à des décisions stratégiques d'investissement faussées (hallucinations, réponses partielles).

Dans la suite, nous traitons les risques à la maille réponse, sauf pour les **Citations défaillantes** et les **Hallucinations** qui sont traitées à la maille phrase. Par ailleurs, nous choisissons de ne pas traiter le problème des **Réponses partielles**, dont l'évaluation nécessite une expertise métier très spécialisée, ni de la **Langue de bois**, qui est une notion relativement subjective et moins critique.

5 Construction du jeu de données

Jeu de questions Nous avons reconstitué le patrimoine documentaire de l'entreprise cible, DataCorp (nom modifié), un cabinet de conseil en informatique, en intégrant 300 documents dans notre base. En parallèle, un questionnaire de due diligence de 121 questions a été élaboré avec des experts métiers, puis classé par thème et par niveau de difficulté (voir Table 1). Certaines questions, inappropriées au contexte, ont été conservées pour évaluer la capacité du système à s'abstenir de répondre [2].

Thème/Difficulté	Simple	Intermédiaire	Difficile	Inapproprié
Finance	16	16	16	15
RH	7	7	7	3
IT	10	10	10	4

TABLE 1 – Répartition des questions de notre jeu de données par thème et par niveau de difficulté.

2. Le jeu de données anonymisé est disponible à cette adresse : https://github.com/gmartinonQM/eval-rag. Le détail de la procédure d'anonymisation est présenté en Section 10.

Jeu de réponses Pour chaque question, nous avons généré 20 réponses distinctes, en conservant le même libellé et en initiant une nouvelle conversation à chaque fois. Cette procédure permet de prendre en compte les **réponses stochastiques** à température non nulle [3].

Extraction de phrases Afin de mieux quantifier et localiser les hallucinations, nous découpons chaque réponse en phrases, comme indiqué en Figure 2.

Stratégie d'échantillonnage Pour les réponses comme pour les phrases, nous sélectionnons les observations à annoter par un plan de sondage stratifié. Dans un premier temps, nous effectuons un embedding de toutes les observations avec le moteur `text-embedding-ada-002` d'OpenAI. Ensuite, nous appliquons un algorithme de clustering K-Means à ces embeddings au sein de chaque thème. Enfin, nous sélectionnons aléatoirement trois observations par cluster. Le nombre de clusters K est déterminé par le budget alloué à l'annotation humaine.

Protocole d'annotation Les annotations, humaines ou produites par un LLM-Juge, permettent de détecter les hors-sujets et les hallucinations. Les consignes, identiques pour les deux types d'annotateurs, sont formalisées sous forme de prompt. Trois annotateurs humains, de niveaux d'expertise croissants, interviennent successivement : le second relit le premier, et un troisième tranche en cas de désaccord persistant. Tous sont coauteurs du présent article. Un exemple d'annotation est présenté en Section 10.

LLM-Juge Dans cet article, le LLM-Juge est GPT-4o (`gpt-4o-2024-08-06`) utilisé à température nulle.

Le jeu de données ainsi constitué ne saurait être considéré comme un *gold standard* : obtenir une réponse idéale à chaque question impliquerait un coût humain bien supérieur. C'est précisément tout l'intérêt de notre approche : permettre une évaluation à bas coût d'un système génératif, tout en s'appuyant sur une revue humaine experte.

6 Protocole d'évaluation

6.1 Métriques

Pour chaque risque identifié, nous appliquons des métriques de performance spécifiques.

Taux de langue correcte (maille réponse) Nous mesurons la proportion de réponses rédigées en français, sachant que toutes les questions sont posées dans cette langue et que nous générons une

3. Le choix de 20 répétitions repose sur une étude exploratoire de répétabilité : nous avons observé que le taux de véracité par question se stabilisait seulement à partir de ce seuil. Un budget plus important aurait permis de générer davantage de répétitions pour obtenir des métriques robustes à l'échelle de chaque question. En pratique, pour limiter les coûts de calcul et d'annotation, nous restituons les résultats agrégés à la maille thème, et considérons que 20 réponses suffisent déjà à capturer une variabilité suffisante du système à cette maille.

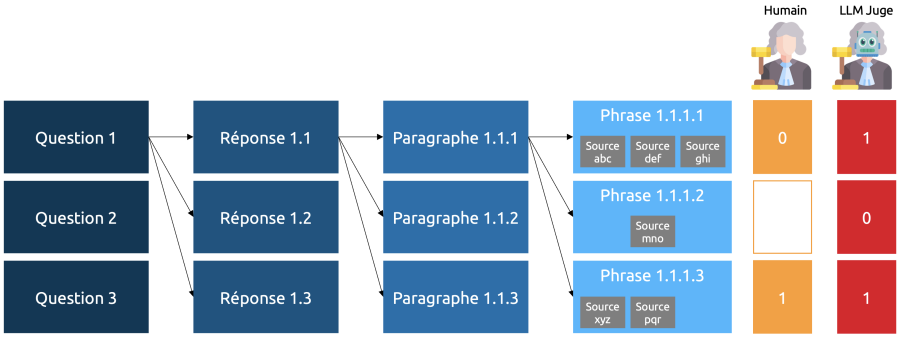

FIGURE 2 – Protocole d'évaluation du système RAG. Pour chaque question, 20 réponses différentes sont générées. Les réponses sont découpées en phrases, chacune avec ses sources. Un annotateur humain et LLM-Juge évaluent les évaluent avec une notation binaire (0 ou 1). Le juge humain n'évalue qu'un échantillon aléatoire, tandis que le LLM-Juge évalue l'intégralité.

vingtaine de réponses par question [4].

Taux de réponse (maille réponse) Nous calculons la proportion de réponses qui ne sont pas des abstentions manifestes et qui citent explicitement des sources. Nous employons des expressions régulières (REGEX) pour détecter la présence de citations dans le texte.

Taux de citations fonctionnelles (maille phrase) Nous calculons la proportion de phrases où tous les IDs de sources citées correspondent aux IDs fournis par le moteur de recherche.

Pertinence (maille réponse) Nous estimons la proportion de réponses exemptes de tout contenu hors sujet. Nous introduisons un label binaire : "0" si la réponse contient au moins un hors sujet, "1" si la réponse ne contient aucun hors sujet.

Véracité (maille phrase) Nous évaluons la proportion de phrases dont toutes les affirmations sont déductibles du texte. Nous introduisons un label binaire : "0" la phrase contient au moins un élément non déductible des sources citées (hallucination), "1" la phrase est entièrement déductible des sources citées. Les phrases qui relèvent de la langue de bois sont annotées "1", pour ne pas introduire de confusion avec la notion d'hallucination.

Les métriques **taux de langue correcte**, **taux de citations fonctionnelles**, et **taux de réponse** peuvent être mesurées de manière programmatique et sans ambiguïté. En revanche, les métriques de **pertinence** et de **véracité** requièrent du raisonnement, que nous traitons par des annotations humaines ou LLM-Juge. Les prompts utilisés à cette fin sont présentés en Section 10.

4. Pour détecter la langue des réponses, nous utilisons le package Python `langdetect`, qui n'a montré aucune erreur sur un échantillon de 1 000 réponses vérifiées manuellement.

6.2 Prediction-Powered Inference

Pour obtenir des intervalles de confiance sur ces métriques, nous appliquons la méthode PPI++ (Angelopoulos *et al.*, 2023b). Cette méthode vise à fournir une estimation fiable de ces métriques en combinant les annotations humaines sur un échantillon sélectionné avec les annotations réalisées par le LLM-Juge sur l'ensemble du jeu de données. Nous donnons plus de détails sur la méthode en Section 10.

7 Résultats

La Figure 3 présente les métriques calculées automatiquement (langue correcte, réponse effective et citation fonctionnelle) par thématique et niveau de difficulté des questions.

On observe une nette dégradation des performances lorsque la complexité des questions augmente. Le système tend alors à s'abstenir de répondre et à basculer en anglais, un comportement également observé pour les questions inappropriées, pourtant censées rester sans réponse. Cela suggère que face à une faible confiance (mesuré par exemple par sa perplexité (Jelinek *et al.*, 1977)), le modèle opte pour l'anglais. Par ailleurs, le système fournit une réponse même lorsqu'aucune n'est attendue. Enfin, bien qu'il restitue correctement les identifiants de sources dans 99% des cas, les erreurs résiduelles peuvent nuire à son exploitation.

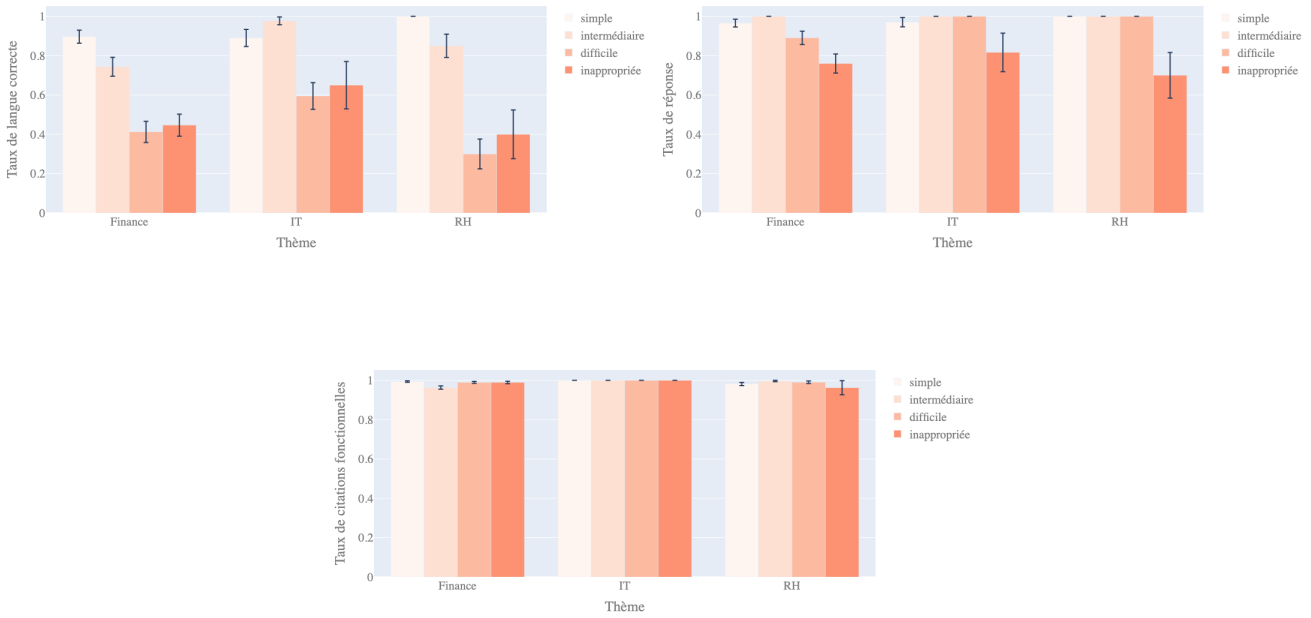

FIGURE 3 – Métriques calculées automatiquement par thème et par niveau de difficulté des questions. En haut à gauche : taux de réponses en français. En haut à droite : taux de réponses effectives. En bas : taux de citations correctes, indiquant si l'identifiant de la source citée est non corrompu. Les barres d'erreurs correspondent aux intervalles de confiance de Wald à 95%.

La Figure 4 présente les métriques nécessitant un raisonnement (pertinence et véracité) et compare les performances des trois méthodes d'évaluation : annotations humaines, LLM-Juge et PPI.

Les annotations humaines révèlent une forte variabilité de pertinence selon les thématiques, avec des résultats particulièrement faibles en IT (32% de réponses entièrement pertinentes). La véracité reste relativement élevée, entre 80% et 88% des phrases. Une analyse manuelle révèle que les 12 à 20% de phrases contenant une hallucination sont bien réparties dans quasiment toutes les réponses.

Comparé à l'humain, le LLM-Juge surestime nettement la pertinence, notamment en IT (écart d'un facteur 2), tandis que les écarts de véracité restent plus modérés (jusqu'à 6%).

Les estimations PPI, très proches de celles obtenues par annotations humaines, montrent que dans cette étude, les annotations du LLM-Juge apportent peu d'information utile.

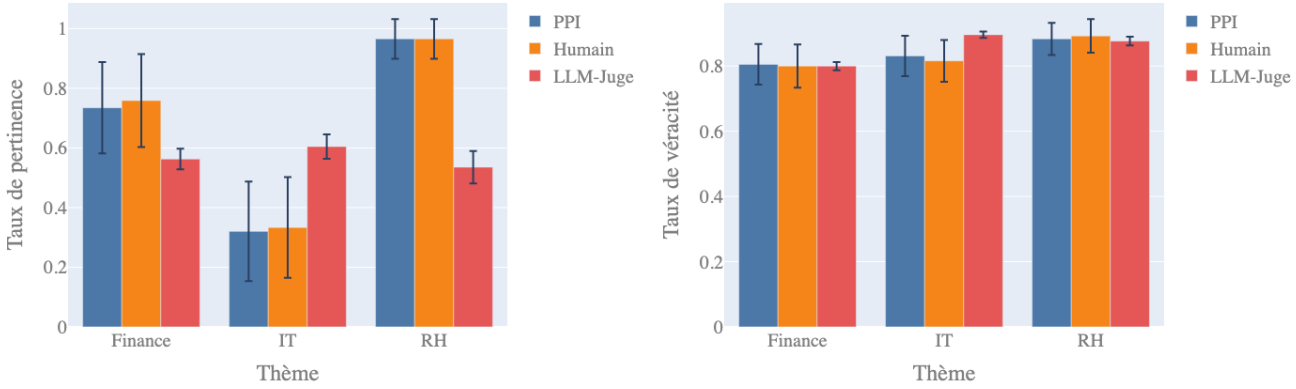

FIGURE 4 – Métriques calculées par annotations humaine et LLM-Juge par thème. A gauche : taux de pertinence. A droite : taux de véracité. Les barres d'erreurs pour "Humain" et "LLM" correspondent aux intervalles de confiance de Wald à 95%, tandis que les barres d'erreurs pour "PPI" sont directement issues de la méthode PPI.

Ce résultat s'explique par le faible taux d'accord entre annotations humaines et LLM-Juge sur l'échantillon de contrôle. Comme le montrent les Tables 2 et 3, l'accord observé frôle parfois le niveau aléatoire, signalant une faible exploitabilité du LLM-Juge. Dans ces conditions, l'incertitude obtenue par PPI est proche de celle d'un simple sondage exploitant les annotations humaines, ce qui rend la contribution du LLM-Juge marginale. Ces constats rejoignent ceux de (Gligorić *et al.*, 2024) sur d'autres jeux publics. Une analyse de sensibilité à l'accord humain/LLM-Juge est proposée en Section 10.

8 Conclusion

Dans cet article, nous avons évalué un système RAG dans un contexte industriel de due diligence, en analysant ses défaillances par domaine métier et en combinant plusieurs méthodes d'évaluation. Nos résultats montrent que, malgré leur potentiel pour automatiser l'analyse documentaire, ces systèmes soulèvent des enjeux critiques de fiabilité, notamment en termes de pertinence et de véracité.

Le protocole proposé, fondé sur des métriques automatiques, des annotations humaines et des évaluations par LLM-Juge, permet une mesure scalable et réplicable à chaque itération du système. Les deux facteurs déterminants pour réduire l'incertitude sont le volume d'annotations humaines et la

Thème	Accord aléatoire	Accord observé	Ann. hum.	Ann. hum. eff.	Ann. LLM-Juge
Finance	0.62	0.69	29	31.98	791
IT	0.43	0.50	30	30.04	551
RH	0.61	0.59	29	29.00	325

TABLE 2 – Gain apporté par la combinaison du LLM-Juge et de PPI pour l'évaluation de la pertinence. L'accord aléatoire correspond au taux attendu si les annotations humaines et LLM-Juge étaient générées indépendamment selon une loi de Bernoulli ; il sert de référence de pire cas. L'accord observé indique la proportion réelle d'annotations concordantes sur l'échantillon de contrôle. Ann. hum. et Ann. LLM-Juge désignent respectivement le volume d'annotations humaines sur l'échantillon et celui, complet, du LLM-Juge. Ann. hum. eff. représente la taille effective des annotations humaines, qui aurait été nécessaire avec un sondage classique pour atteindre la même incertitude que celle observée avec PPI.

Thème	Accord aléatoire	Accord observé	Ann. hum.	Ann. hum. eff.	Ann. LLM-Juge
Finance	0.67	0.79	140	155.59	3985
IT	0.73	0.80	141	141.57	3799
RH	0.82	0.88	139	162.36	2408

TABLE 3 – Gain apporté par l'utilisation conjointe du LLM-Juge et de PPI pour l'évaluation de la véracité.

qualité du LLM-Juge. Or, la méthode PPI ne devient réellement avantageuse que si l'accord entre humain et LLM-Juge est excellent, ce qui suppose un travail rigoureux de conception de prompt, parfois plus coûteux en temps qu'une session supplémentaire d'annotation humaine.

Intégrée dès les premières phases de développement, cette approche permet néammoins d'alimenter une boucle de rétroaction continue : affiner le prompt du LLM-Juge, ajuster les consignes d'annotation ou enrichir un corpus pour du *fine-tuning* du LLM-Juge. Ce cycle de capitalisation progressive constitue, selon nous, une voie prometteuse pour fiabiliser durablement l'évaluation des systèmes génératifs.

9 Limitations et directions futures

Le protocole proposé ouvre plusieurs perspectives d'amélioration.

D'abord, la granularité d'annotation pourrait être affinée : au lieu d'évaluer la véracité au niveau des phrases, on pourrait cibler des plages de mots, comme dans les jeux de données ANAH (Ji *et al.*, 2024) ou HaluEval (Li *et al.*, 2023).

Ensuite, l'échantillonnage utilisé ici repose sur des embeddings généralistes. L'adoption d'un moteur d'embedding adapté au contexte de la due diligence pourrait produire un clustering plus pertinent, et donc réduire l'incertitude via un échantillonnage mieux stratifié.

Notre étude s'est par ailleurs limitée à un seul LLM-Juge (GPT-4o). En comparant plusieurs modèles et prompts, il serait possible d'identifier les configurations les plus fiables, en s'appuyant sur l'incertitude calculée par PPI comme indicateur de performance. Une telle sélection de couple LLM/prompt nécessiterait toutefois un jeu d'annotations de validation pour éviter tout sur-apprentissage.

Le protocole pourrait aussi être étendu à des variantes méthodologiques, telles que ASI, mieux

adaptées aux sondages stratifiés que ne l'est PPI, fondé sur un plan de sondage simple.

Enfin, une piste supplémentaire consisterait à intégrer la langue de bois comme nouvelle dimension d'évaluation, à l'aide d'un protocole dédié.

N.B. Cet article a été rédigé avec l'assistance de GPT-4o.

10 Matériel supplémentaire

10.1 Procédure d'anonymisation des données

Les données ont été systématiquement anonymisées selon le protocole suivant :

- Remplacement de toutes les adresses postales et noms de ville
- Substitution de tous les prénoms et noms de famille
- Changement des noms d'entreprise
- Réécriture du vocabulaire spécifique à l'entreprise cible
- Remplacement des adresses e-mail par `johndoe@company.com`
- Décalage temporel de toutes les dates vers le passé
- Substitution des liens web par `https://example.com`
- Modification de tous les nombres, pourcentages et montants financiers
- Remplacement des numéros de téléphone par `01 23 45 67 89`
- Remplacement des numéros SIREN par `123 456 789`
- Remplacement des numéros de compte bancaire par `123456789`

Ce protocole d'anonymisation vise à garantir la protection de l'identité de l'entreprise cible, de l'entreprise acquérante ainsi que de toutes les personnes mentionnées dans les documents analysés.

10.2 Exemples d'annotations

La figure 5 illustre une capture d'écran d'un fichier excel ayant permis d'annoter une phrase générée par le système évalué Alban. En l'occurrence, l'exemple illustre une hallucination au sens où la réponse et la source se situent à des dates différentes.

10.3 Prompts du LLM-Juge

Le prompt du LLM-Juge sur la véracité est présenté en Figure 6.

Le prompt du LLM-Juge sur la pertinence est présenté en Figure 7.

10.4 PPI

En notant n la taille de l'échantillon annoté par l'homme, X_j les réponses associées, Y_j les annotations humaines correspondantes, N la taille de l'échantillon uniquement annoté par le LLM-Juge, $\tilde{X}_i$ les réponses associées, et $f(.)$ l'annotation LLM-Juge, on peut estimer de manière non biaisée la moyenne des Y_i sur l'ensemble des données, notée $\hat{\theta}$, par l'expression suivante :

Concernant les indicateurs de 2017. Quel fut le taux de TACE en 2017 ? En fin d'année, le TACE était de 69,8%, avec un pic en janvier à 76,6%. Le taux moyen étant de 66,5%.		
Statements	**Déductible (1= Oui, 0= Non)**	**Réponse**
Le taux moyen était de 66,5%[^4b8b6ff358bf79e7c95c24bd4894421209dd79df^].	0	DataCorp utilise plusieurs indicateurs de performance pour ses plans d'incitation annuels. Taux d'Activité Commerciale Effective (TACE): Le TACE est un autre indicateur de performance utilisé par DataCorp. En 2019, le TACE était de 69,8%, avec un pic en janvier à 76,6%. Le taux moyen était de 66,5%[^4b8b6ff358bf79e7c95c24bd4894421209dd79df^].

FIGURE 5 – Exemple d'annotation d'une hallucination. L'énoncé présenté ("Le taux moyen était de 66,5%") est incorrectement déduit de la source (en haut à gauche) : la source décrit l'année 2017, quand la réponse décrit l'année 2019. L'annotation attribue donc le label **0** (non déductible).

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} \lambda f(\tilde{X}_i) - \frac{1}{n}\sum_{j=1}^{n} (\lambda f(X_j) - Y_j) \tag{1}$$

En première approche, le paramètre λ peut être considéré égal à 1. Cependant, comme indiqué dans (Angelopoulos *et al.*, 2023b), si les évaluations du LLM-Juge sont de mauvaise qualité, les inclure dans l'estimation peut élargir l'intervalle de confiance par rapport à ce qui serait obtenu avec un sondage simple basé uniquement sur les Y_j. Pour éviter ce problème, on introduit le paramètre $\lambda \in [0, 1]$, qui peut être optimisé pour garantir des estimations aussi précises, voire meilleures, que celles obtenues par un sondage simple, indépendamment de la qualité des annotations du LLM-Juge ; c'est le *power tuning*. L'expression analytique du paramètre λ optimal dépend des observations faites sur le jeu de données.

L'intervalle de confiance au niveau $1 - \alpha$ sur $\hat{\theta}$ s'obtient, lorsque $\lambda = 1$, par :

$$C_{1-\alpha} = \hat{\theta} \pm z_{1-\alpha/2}\sqrt{\left(\frac{\hat{\sigma}_f^2}{N} + \frac{\hat{\sigma}_{f-Y}^2}{n}\right)} \tag{2}$$

où $z_{1-\alpha/2}$ est le quantile $1-\alpha/2$ de la distribution normale standardisée, $\hat{\sigma}_f^2$ et $\hat{\sigma}_{f-Y}^2$ représentent les estimations de variance de $f(\tilde{X}_i)$ et $f(X_j) - Y_j$ respectivement. La formule analytique dans le cas où $\lambda \neq 1$ est donnée en Section 6 de (Angelopoulos *et al.*, 2023b). Dans cet article, nous utilisons le package python `ppi_py` disponible à cette adresse `https://github.com/aangelopoulos/ppi_py`, et développé par les auteurs de la méthode.

10.5 Analyse de sensibilité de PPI à l'accord humain/LLM-Juge

Ayant pu observer que le gain sur l'incertitude de mesure apporté par le LLM-Juge et PPI était marginal, notamment à cause d'un taux d'accord insuffisant entre les annotations humaines et LLM-

Rôle du LLM-Juge :
Votre mission est d'analyser la véracité des phrases fournies par rapport à des documents de référence. Dès qu'une information de la phrase est non-déductible des sources, alors toute la phrase doit être classée comme non-déductible.

Les données qui vous seront fournies sont :

— **Sources** : Les documents de référence.
— **Paragraphe** : Le paragraphe contextualisant la phrase.
— **Phrases** : La phrase dont vous devez évaluer le caractère déductible.

Instructions d'évaluation :
Adoptez l'approche suivante dans votre évaluation :

1. Lire les sources en entier.
2. Lire le paragraphe en entier.
3. Lire les phrases en entier.
4. Pour chaque phrase, verbaliser un raisonnement pas à pas sur le caractère déductible des informations de la phrase par rapport aux sources.
5. Si les informations de la phrase sont insuffisantes ou ambiguës, vous pouvez utiliser vos connaissances du monde pour déterminer si toutes ces informations sont réellement déductibles des sources.
6. En déduire une évaluation finale pour chaque phrase.

Chaque évaluation est un label :

— **0** : Au moins une information n'est pas déductible des sources.
— **1** : Toutes les informations sont déductibles des sources.

Format des réponses :
Vous devez renvoyer uniquement un JSON dans la structure suivante : *json schema*
Important : Le contenu renvoyé doit être un JSON strictement valide, sans texte supplémentaire, sans explication ni commentaire, directement parsable et la clé « verdicts » devant correspondre à une liste d'objets JSON.

Exemples :
Exemple de sources : *example input*
Exemple de paragraphes : *example paragraph*
Exemple de phrase : *example statements*
Exemple de réponse au format JSON : : *example output*

Voici les données :
Sources : *sources*
Paragraphes : *paragraphes*
Phrases : *phrases*
JSON :

FIGURE 6 – Prompt du LLM-Juge pour la véracité. Le LLM lit les sources et les phrases, applique un raisonnement étape par étape, puis attribue un label (déductible, non-déductible) en suivant un protocole strict. Le résultat final est structuré sous forme d'un fichier JSON.

Rôle du LLM-Juge :
Votre mission est d'évaluer la pertinence d'une réponse par rapport à une question posée. Une réponse est dite pertinente si elle :

- répond à la question posée,
- n'introduit pas d'éléments hors sujet,

Il ne s'agit pas ici d'évaluer la véracité de la réponse, ni son objectivité.

Les données qui vous seront fournies sont :

- **Question** : La question posée.
- **Réponse** : La réponse à évaluer.

Instructions d'évaluation :
Adoptez l'approche suivante dans votre évaluation :

1. Lire attentivement la question.
2. Lire attentivement la réponse.
3. Lire les phrases en entier.
4. Pour chaque phrase, verbaliser un raisonnement pas à pas sur le caractère pertinent des informations de la réponse.
5. En déduire une évaluation finale pour chaque phrase.

Chaque évaluation est un label :

- **0** : La réponse est partiellement pertinente et contient au moins un hors sujet.
- **1** : La réponse est totalement pertinente.

Format des réponses :
Vous devez renvoyer uniquement un JSON dans la structure suivante : *json schema*
Important : Le contenu renvoyé doit être un JSON strictement valide, sans texte supplémentaire, sans explication ni commentaire, directement parsable et la clé « verdicts » devant correspondre à une liste d'objets JSON.

Exemples :
Voici un ensemble d'exemples de phrases et de leurs évaluations.
Exemple de question : *example question*
Exemple de réponse : *example answer*
Exemple de réponse au format JSON : : *example output*

Voici les données :
Question : *question*
Réponse : *answer*
JSON :

FIGURE 7 – Prompt du LLM-Juge pour la pertinence. Le LLM lit la question et la réponse, applique un raisonnement étape par étape, puis attribue un label (totalement pertinent, contient au moins un hors-sujet) en suivant un protocole strict. Le résultat final est structuré sous forme d'un fichier JSON.

Juge sur l'échantillon de contrôle, on peut légitimement se demander comment évolue l'incertitude de PPI avec l'accord humain/LLM-Juge.

Nous avons mené une simulation reproduisant les caractéristiques observées sur la métrique de véracité en finance (cf. Table 3), tout en faisant varier artificiellement le taux d'accord entre annotations humaines et LLM-Juge. Dans cette simulation, nous avons considéré 140 annotations humaines, 3985 annotations LLM-Juge, et des métriques mesurées par l'homme et le LLM-Juge identiques, toutes deux égales à 0.8. Le code de simulation est disponible en libre accès : https://github.com/gmartinonQM/eval-rag.

On constate que pour passer d'une incertitude de 7% (sondage classique avec 140 observations) à une incertitude de 4%, il faut être capable de concevoir un LLM-Juge qui s'accorde avec l'être humain dans 93% des cas. Si tel était le cas, l'incertitude de mesure de 4% ainsi obtenue serait la même qu'un sondage classique utilisant 375 annotations, soit un facteur 2.7 de gagné sur le temps d'annotations humaines. Ces résultats illustrent de manière quantitative à quel point la qualité du LLM-Juge est déterminante pour que la méthode PPI apporte un avantage substantiel en termes de coût et d'efficacité.

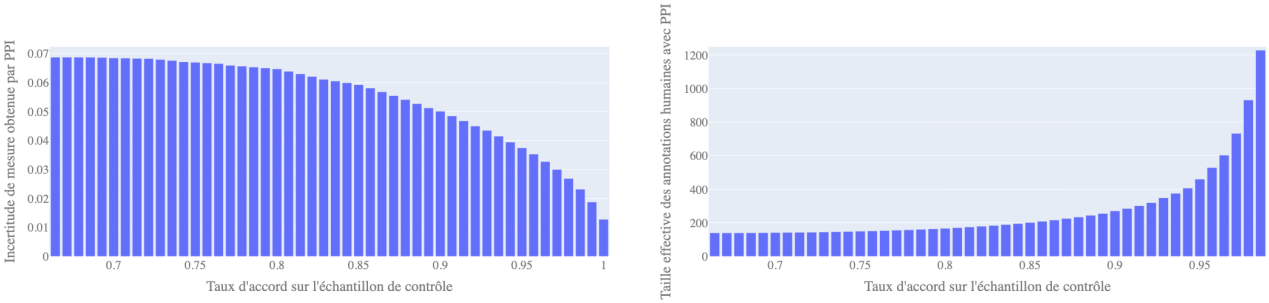

FIGURE 8 – Simulation de résultats obtenus par PPI en faisant varier l'accord humain/LLM-Juge sur l'échantillon de contrôle. A gauche : incertitude de mesure. A droite : taille effective des annotations humaine.

Références

ANGELOPOULOS A. N., BATES S., FANNJIANG C., JORDAN M. I. & ZRNIC T. (2023a). Prediction-powered inference. *Science*, **382**(6671), 669–674. DOI : 10.1126/science.adi6000.

ANGELOPOULOS A. N., DUCHI J. C. & ZRNIC T. (2023b). Ppi++ : Efficient prediction-powered inference. *arXiv preprint arXiv :2311.01453*.

BLANDÓN M. A. C., TALUR J., CHARRON B., LIU D., MANSOUR S. & FEDERICO M. (2025). Memerag : A multilingual end-to-end meta-evaluation benchmark for retrieval augmented generation. *arXiv preprint arXiv :2502.17163*.

BOYEAU P., ANGELOPOULOS A. N., YOSEF N., MALIK J. & JORDAN M. I. (2024). Autoeval done right : Using synthetic data for model evaluation. *arXiv preprint arXiv :2403.07008*.

CHERN I., CHERN S., CHEN S., YUAN W., FENG K., ZHOU C., HE J., NEUBIG G., LIU P. *et al.* (2023). Factool : Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv :2307.13528*.

CHERN S., CHERN E., NEUBIG G. & LIU P. (2024). Towards scalable oversight : Meta-evaluation of LLMs as evaluators via agent debate. In *2nd AI4Research Workshop : Towards a Knowledge-grounded Scientific Research Lifecycle*.

ES S., JAMES J., ANKE L. E. & SCHOCKAERT S. (2024). Ragas : Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, p. 150–158.

FARQUHAR S., KOSSEN J., KUHN L. & GAL Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, **630**(8017), 625–630.

FRIEL R. & SANYAL A. (2023). Chainpoll : A high efficacy method for llm hallucination detection. *arXiv preprint arXiv :2310.18344*.

GAO Y., XIONG Y., GAO X., JIA K., PAN J., BI Y., DAI Y., SUN J., WANG H. & WANG H. (2023). Retrieval-augmented generation for large language models : A survey. *arXiv preprint arXiv :2312.10997*, **2**.

GLIGORIĆ K., ZRNIC T., LEE C., CANDÈS E. J. & JURAFSKY D. (2024). Can unconfident llm annotations be used for confident conclusions ? *arXiv preprint arXiv :2408.15204*.

HONG G., GEMA A. P., SAXENA R., DU X., NIE P., ZHAO Y., PEREZ-BELTRACHINI L., RYABININ M., HE X., FOURRIER C. *et al.* (2024). The hallucinations leaderboard–an open effort to measure hallucinations in large language models. *arXiv preprint arXiv :2404.05904*.

JELINEK F., MERCER R. L., BAHL L. R. & BAKER J. M. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, **62**.

JI Z., GU Y., ZHANG W., LYU C., LIN D. & CHEN K. (2024). ANAH : Analytical annotation of hallucinations in large language models. 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. 8135–8158, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.442.

JUNG J., BRAHMAN F. & CHOI Y. (2025). Trust or escalate : LLM judges with provable guarantees for human agreement. In *The Thirteenth International Conference on Learning Representations*.

LI J., CHENG X., ZHAO X., NIE J.-Y. & WEN J.-R. (2023). HaluEval : A large-scale hallucination evaluation benchmark for large language models. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 6449–6464, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.397.

LIANG P., BOMMASANI R., LEE T., TSIPRAS D., SOYLU D., YASUNAGA M., ZHANG Y., NARAYANAN D., WU Y., KUMAR A., NEWMAN B., YUAN B., YAN B., ZHANG C., COSGROVE C., MANNING C. D., RE C., ACOSTA-NAVAS D., HUDSON D. A., ZELIKMAN E., DURMUS E., LADHAK F., RONG F., REN H., YAO H., WANG J., SANTHANAM K., ORR L., ZHENG L., YUKSEKGONUL M., SUZGUN M., KIM N., GUHA N., CHATTERJI N. S., KHATTAB O., HENDERSON P., HUANG Q., CHI R. A., XIE S. M., SANTURKAR S., GANGULI S., HASHIMOTO T., ICARD T., ZHANG T., CHAUDHARY V., WANG W., LI X., MAI Y., ZHANG Y. & KOREEDA Y. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification, Outstanding Certification.

LIU Y., ITER D., XU Y., WANG S., XU R. & ZHU C. (2023). G-eval : NLG evaluation using gpt-4 with better human alignment. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 2511–2522, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.153.

MAGESH V., SURANI F., DAHL M., SUZGUN M., MANNING C. D. & HO D. E. (2024). Hallucination-free ? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv :2405.20362*.

MALEKI N., PADMANABHAN B. & DUTTA K. (2024). Ai hallucinations : a misnomer worth clarifying. In *2024 IEEE conference on artificial intelligence (CAI)*, p. 133–138 : IEEE.

MIN S., KRISHNA K., LYU X., LEWIS M., TAU YIH W., KOH P. W., IYYER M., ZETTLEMOYER L. & HAJISHIRZI H. (2023). FActscore : Fine-grained atomic evaluation of factual precision in long form text generation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

MOUNTANTONAKIS M. & TZITZIKAS Y. (2023). Validating chatgpt facts through rdf knowledge graphs and sentence similarity. *arXiv preprint arXiv :2311.04524*.

NGUYEN T., ROSENBERG M., SONG X., GAO J., TIWARY S., MAJUMDER R. & DENG L. (2016). Ms marco : A human generated machine reading comprehension dataset. In T. R. BESOLD, A. BORDES, A. S. D'AVILA GARCEZ & G. WAYNE, Éds., *CoCo@NIPS*, volume 1773 de *CEUR Workshop Proceedings* : CEUR-WS.org.

NIU C., WU Y., ZHU J., XU S., SHUM K., ZHONG R., SONG J. & ZHANG T. (2024). RAGTruth : A hallucination corpus for developing trustworthy retrieval-augmented language models. 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. 10862–10878, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.585.

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

SCIRÈ A., BEJGU A. S., TEDESCHI S., GHONIM K., MARTELLI F. & NAVIGLI R. (2024). Truth or mirage ? towards end-to-end factuality evaluation with llm-oasis. *arXiv preprint arXiv :2411.19655*.

TANG L., SHALYMINOV I., WONG A., BURNSKY J., VINCENT J., YANG Y., SINGH S., FENG S., SONG H., SU H., SUN L., ZHANG Y., MANSOUR S. & MCKEOWN K. (2024). TofuEval : Evaluating hallucinations of LLMs on topic-focused dialogue summarization. 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. 4455–4480, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.251.

VALENTIN S., FU J., DETOMMASO G., XU S., ZAPPELLA G. & WANG B. (2024). Cost-effective hallucination detection for llms. *arXiv preprint arXiv :2407.21424*.

WEI J., YANG C., SONG X., LU Y., HU N., HUANG J., TRAN D., PENG D., LIU R., HUANG D., DU C. & LE Q. V. (2024). Long-form factuality in large language models. In A. GLOBERSON, L. MACKEY, D. BELGRAVE, A. FAN, U. PAQUET, J. TOMCZAK & C. ZHANG, Éds., *Advances in Neural Information Processing Systems*, volume 37, p. 80756–80827 : Curran Associates, Inc.

WEIDINGER L., RAJI D., WALLACH H., MITCHELL M., WANG A., SALAUDEEN O., BOMMASANI R., KAPOOR S., GANGULI D., KOYEJO S. *et al.* (2025). Toward an evaluation science for generative ai systems. *arXiv preprint arXiv :2503.05336*.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E. *et al.* (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, **36**, 46595–46623.

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

ZRNIC T. & CANDÈS E. J. (2024). Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24 : JMLR.org.