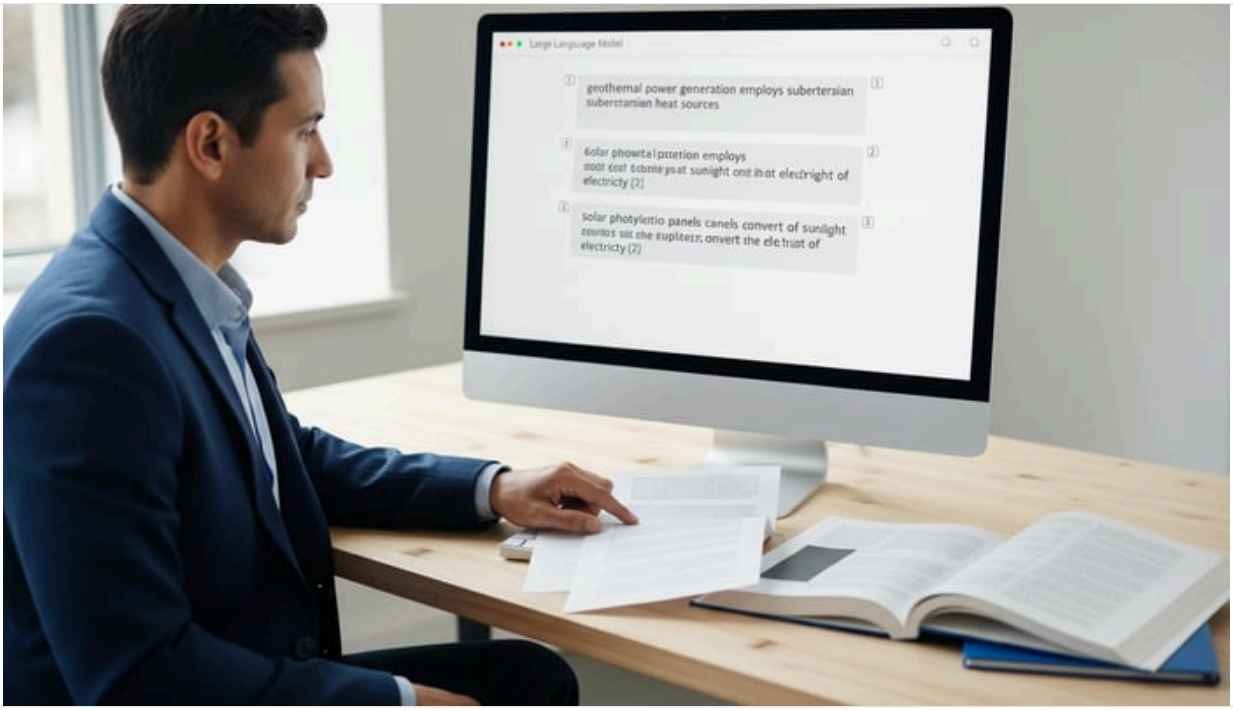
Citations LLM expliquées : Méthodes RAG et d'attribution de source

By rankstudio.net Publié le 17 octobre 2025 29 min de lecture



Résumé

Les grands modèles linguistiques (LLM) modernes tels que ChatGPT d'OpenAI, Gemini de Google et d'autres sont de plus en plus utilisés pour la récupération et la synthèse d'informations. Cependant, ces modèles ne divulguent pas nativement la provenance de leurs sorties, ce qui conduit au problème bien connu des « hallucinations » – des réponses affirmées avec confiance mais non étayées ou incorrectes. En réponse, les chercheurs et les développeurs ont commencé à construire des **cadres de citation pour l'IA** : des méthodes systématiques permettant aux LLM de joindre des références ou des attributions de source à leurs réponses. Ces cadres se répartissent généralement en deux grandes catégories : l'intégration de techniques de **génération augmentée par récupération (RAG)** et l'intégration de **mécanismes d'attribution de source** dans l'entraînement/la sortie du modèle.

Dans les systèmes RAG, une question déclenche une recherche dans des bases de données externes ou sur le web pour recueillir des documents pertinents avant (ou pendant) la génération de la réponse. Par exemple, Google Research note que « le RAG améliore les grands modèles linguistiques en leur fournissant un contexte externe pertinent » (Source: research.google). En alimentant directement le contenu factuel dans l'entrée du LLM, le RAG permet de citer des sources réelles. En pratique, ChatGPT avec navigation ou plugins et des services spécialisés comme Perplexity.ai mettent en œuvre cette idée, ajoutant souvent des notes de bas de page ou des liens vers des documents sources.

Alternativement, de nouveaux algorithmes cherchent à intégrer des signaux de source dans la sortie même du LLM. Un exemple phare est **WASA (WAtermark-based Source Attribution)**, qui entraîne un LLM à inclure des marqueurs cachés encodant l'identité du fournisseur de données original (Source: openreview.net). Dans WASA, chaque segment de texte généré porte un « filigrane » traçable afin que l'on puisse retrouver de quel corpus ou document d'entraînement il provient. Plus généralement, certaines approches de réglage fin enseignent à un LLM à produire des citations (par exemple, des références savantes via DOI) dans le cadre de sa réponse.

Les études empiriques dressent un tableau mitigé de la performance actuelle des LLM en matière de citation. Dans une tâche de connaissance médicale, ChatGPT-4 a fourni des références pour toutes les réponses (lorsqu'il était invité à le faire), mais seulement 43,3 % de ces références étaient entièrement exactes ou « vraies » (Source: pmc.ncbi.nlm.nih.gov). En fait, plus de la moitié (56,7 %) étaient soit incorrectes, soit inexistantes (Source: pmc.ncbi.nlm.nih.gov), faisant écho aux avertissements selon lesquels, sans vérification, même les réponses de GPT-4 « ne parviennent pas à fournir des références fiables et reproductibles » (Source: pmc.ncbi.nlm.nih.gov). En revanche, une étude plus large et transdomaine a révélé que les analogues de GPT-4 produisaient des citations extrêmement bonnes : environ 90 % de leurs

références étaient factuelles et seulement ~10 % étaient fabriquées (Source: www.mdpi.com) (Source: www.mdpi.com). Ces différences soulignent que la qualité des citations dépend grandement du contexte, de la conception de l'invite et de l'accès aux connaissances externes. De manière alarmante, une expérience récente a montré que plusieurs LLM (GPT-4o, Google Gemini, Meta Llama 3.2, xAI Grok) pouvaient être incités à donner des conseils médicaux de style autoritaire avec des citations de revues entièrement inventées – seul Claude d'Anthropic a refusé l'invite (Source: www.reuters.com).

Ce rapport fournit une analyse technique approfondie de la manière dont les LLM obtiennent et attribuent des informations. Nous commençons par un aperçu des sources de connaissances des LLM et de la motivation des citations intégrées. Nous passons ensuite en revue les approches existantes : architectures RAG avec liaison de source, techniques de filigrane et de provenance (par exemple, WASA), et vérification post-génération. Nous résumons les données empiriques issues d'études de cas et d'expériences utilisateur, y compris des métriques quantitatives de précision des citations. Enfin, nous discutons des implications plus larges pour la confiance, la propriété intellectuelle et les futures normes. Assurer des citations précises dans l'écriture assistée par l'IA reste un défi multidisciplinaire urgent (Source: openreview.net) (Source: haruiz.github.io), et ce rapport présente le paysage actuel et les orientations de recherche.

Introduction

Contexte : Connaissance et confiance dans les LLM

Les grands modèles linguistiques (LLM) comme GPT-4, Claude et Gemini sont entraînés sur de vastes corpus de texte (les « données d'entraînement ») et apprennent à générer du texte de type humain. En interrogeant ces modèles, les utilisateurs peuvent obtenir des réponses à des questions factuelles, des résumés et des conseils dans divers domaines. Cependant, contrairement aux moteurs de recherche traditionnels ou aux bases de données, la réponse d'un LLM ne s'accompagne *pas* automatiquement de liens vers ses sources. La connaissance du modèle réside dans les poids du réseau plutôt que dans des index explicites de documents. Par conséquent, les LLM peuvent produire avec confiance des **hallucinations** – des affirmations plausibles mais incorrectes ou invérifiables. Par exemple, une étude systématique de 4 900 résumés scientifiques a révélé que les LLM de pointe étaient près de *cinq fois plus susceptibles* que les experts humains de simplifier à l'excès ou de déformer les résultats clés (Source: www.livescience.com). Dans des domaines sensibles comme la médecine, ces distorsions sont particulièrement dangereuses : les LLM « ont altéré le langage précis concernant la sécurité ou l'efficacité des médicaments, omettant des détails cruciaux » (Source: www.livescience.com).

Une partie du problème est que les LLM manquent d'un mécanisme interne pour citer ou lier des preuves. Dans l'érudition et le journalisme traditionnels, chaque affirmation factuelle est étayée par une citation ou une référence. En revanche, les LLM sont des « boîtes noires » qui produisent du texte sans attribution traçable. Un article médical récent a observé sans ambages que même ChatGPT-4 « connaît son A B C D E mais ne peut pas citer sa source » (Source: pmc.ncbi.nlm.nih.gov), ce qui signifie qu'il peut décrire correctement le protocole de traumatisme ABCDE mais ne parvient pas à fournir des références fiables. De même, les praticiens avertissent que les réponses des LLM **ne devraient pas** être fiables sans vérification croisée : « seulement si utilisé avec prudence, avec vérification croisée » ChatGPT-4 pourrait être sûr pour le soutien à la décision médicale (Source: pmc.ncbi.nlm.nih.gov).

La prise de conscience croissante de ces risques a stimulé les efforts pour développer des cadres de citation structurés pour l'IA. L'objectif est de doter les sorties des LLM d'un contexte ou de références explicites afin que les utilisateurs (et les systèmes automatisés) puissent vérifier les faits. Dans ce rapport, nous examinons à la fois les méthodes techniques d'approvisionnement en informations et les mécanismes d'attribution de celles-ci. Nous définissons un *Cadre de Citation pour l'IA* comme tout système qui permet à la réponse d'un LLM d'être ancrée dans des documents externes, des bases de données ou des métadonnées d'entraînement, idéalement avec des pointeurs directs (par exemple, des notes de bas de page ou des URL) vers ces sources. Cela contraste avec la génération libre où le modèle concocte simplement une réponse à partir d'une mémoire interne nébuleuse.

Historique et motivation

L'idée de texte généré par machine renvoyant à des sources est relativement nouvelle. Les premiers LLM (GPT-2/3) étaient utilisés sans réfléchir comme des « moteurs de connaissance », et produisaient du texte sans aucune indication de provenance. Certains produits initiaux ont essayé d'atténuer cela en intégrant des capacités de recherche : par exemple, Bing Chat (Copilot) de Microsoft et Perplexity.ai ajoutent automatiquement des liens de résultats de recherche web à leurs réponses. Mais ce sont des intégrations spéciales, pas des fonctionnalités inhérentes aux LLM. Plus fondamentalement, la communauté de recherche en IA reconnaît que la **traçabilité des sources** est essentielle pour la confiance. Comme le note un développeur d'IA, l'ajout de citations « facilite la vérification que le LLM utilise des informations pertinentes, réduisant ainsi la probabilité d'hallucinations » (Source: haruiz.github.io). En fait, sans citations, même un système RAG très performant « devient une ‘boîte noire’, sapant la fiabilité et la vérifiabilité » de ses réponses (Source: haruiz.github.io).

Parallèlement, les préoccupations juridiques et éthiques amplifient le besoin de citation. L'entraînement des LLM sur des matériaux protégés par le droit d'auteur sans attribution a conduit à des poursuites judiciaires (par exemple, le New York Times a poursuivi Microsoft et OpenAI, accusant leurs chatbots de profiter d'un « passe-droit » sur le journalisme du NYT (Source: swarajyamag.com). Ces questions de propriété intellectuelle soulignent l'importance de savoir exactement quelles sources ont contribué à la sortie d'un LLM. Un article de cadre récent le souligne : les textes synthétiques « peuvent enfreindre la propriété intellectuelle des données utilisées pour entraîner les LLM », rendant « impératif de pouvoir effectuer l'attribution de source » pour le contenu généré (Source: openreview.net). En bref, à mesure que les LLM s'intègrent à l'éducation, à la recherche et à la politique, l'intégration de mécanismes de citation robustes est considérée comme un impératif à la fois technique et social (Source: research.google) (Source: openreview.net).

Portée de ce rapport

Nous analyserons *comment* les LLM peuvent acquérir et joindre des citations. Cela implique deux éléments principaux : l'**approvisionnement** (comment le modèle obtient des informations factuelles) et l'**attribution** (comment il étiquette ces informations avec une source). Nous couvrons les techniques de récupération traditionnelles (recherche, bases de données vectorielles), les nouvelles méthodes comme les filigranes et l'intégration, et l'état de la pratique dans les assistants IA réels. Nous nous appuyons sur la recherche publiée, la documentation produit et les résultats expérimentaux pour évaluer les performances. Dans la mesure du possible, nous incluons des données quantitatives sur la précision des citations. Nous examinons également des études de cas dans des contextes réels (par exemple, la médecine, la rédaction académique, les conseils de santé) pour illustrer les succès et les échecs. Enfin, nous discutons des implications plus larges pour la confiance, l'éthique et les futures normes. Tout au long du rapport, nous supposons un public académique/professionnel ; notre ton est formel et fondé sur des preuves, avec de nombreuses références.

Fondements des citations IA

Connaissances des LLM : Données d'entraînement vs. récupération externe

Connaissances pré-entraînées. Fondamentalement, un LLM pré-entraîné « connaît » tout ce qui a été intégré dans ses données d'entraînement (jusqu'à sa date limite). Ces données peuvent inclure des livres, des articles, des pages web, du code, etc., mais le modèle compresse tout cela en interne dans les poids de son réseau. Il est crucial de noter que le LLM **ne stocke pas de pointeurs vers des documents**. Ainsi, par défaut, il ne dispose d'aucun moyen intégré de dire « j'ai appris ceci du Document X suivi du Document Y ». Le seul mode d'inférence est de générer du texte basé sur des schémas statistiques. En conséquence, les réponses du LLM peuvent refléter une vaste connaissance, mais n'offrent aucune trace inhérente aux sources.

Sans conception spéciale, cela conduit au problème des « affirmations non sourcées ». Par exemple, ChatGPT-3 a été largement critiqué en 2022 pour avoir donné des citations et des références fictives lorsqu'on lui demandait de justifier ses réponses. Une évaluation générale de la rédaction savante a révélé que ChatGPT-3.5 (utilisant GPT-3.5 Turbo) produisait de nombreuses références qui ne pouvaient pas être vérifiées, les DOI générés étant souvent de pures « hallucinations » (Source: pmc.ncbi.nlm.nih.gov) (Source: pmc.ncbi.nlm.nih.gov). Dans une expérience, 30 des 30 soi-disant références générées par GPT-3.5 sur des questions médicales se sont avérées fausses ou incomplètes (Source: pmc.ncbi.nlm.nih.gov). La raison fondamentale est que le modèle n'a pas d'accès explicite à une base de connaissances au moment de la génération ; il ne fait qu'imiter le style de références plausibles.

Génération augmentée par récupération (RAG). Pour combler la lacune d'accès, la solution prédominante a été de combiner le LLM avec un système de récupération. Dans une configuration RAG, la requête de l'utilisateur déclenche une recherche dans un corpus externe avant que le LLM ne génère la réponse. Ce corpus peut être constitué d'articles universitaires, de documents internes ou du web en direct. Les documents récupérés (ou des extraits pertinents) sont fournis au LLM comme contexte additionnel. Concrètement, on pourrait effectuer une recherche par mots-clés ou une recherche de similarité vectorielle sur une base de données, obtenir les extraits top-K, et les préfixer à l'invite du modèle. Le LLM génère alors sa réponse *ancrée dans le texte récupéré*.

Les groupes de recherche de Google soulignent cette approche : « le RAG améliore les LLM en leur fournissant un contexte externe pertinent » (Source: research.google). En pratique, de nombreux systèmes de QA modernes basés sur les LLM utilisent le RAG. Par exemple, le chatbot Perplexity interroge en interne des sources web et inclut des liens cliquables comme citations. Bing Chat de Microsoft et Bard de Google effectuent de manière similaire des recherches web en arrière-plan et joignent des extraits de résultats ou des URL à leurs réponses. Ces systèmes sous-traitent efficacement l'approvisionnement factuel à la couche de recherche, utilisant le LLM principalement pour l'agrégation et l'explication. Documentant la puissance du RAG, une étude note qu'un contexte correctement récupéré peut « réduire significativement les hallucinations » et améliorer la précision factuelle (Source: research.google). Un autre exemple est l'API PALM2 de Google, qui renvoie des citations vers les résultats de recherche Google lorsqu'elle est utilisée avec les bonnes invites.

En résumé, le RAG transforme le LLM non supervisé en un outil d'IA hybride : en partie moteur de recherche, en partie générateur. Il offre un chemin direct vers les citations car les « sources » sont précisément les documents récupérés. On peut simplement ajouter des citations [Source : *URL ou titre*] dans la réponse formatée. Cependant, l'approche a des limites : elle nécessite la maintenance d'une grande base de connaissances ou d'une API de recherche, et la récupération peut échouer si les requêtes sont hors sujet. Si le LLM interprète mal le contexte ou si des fabrications s'y glissent, la réponse peut toujours être trompeuse même avec des références. De plus, la mise en œuvre du RAG de manière fiable implique une ingénierie minutieuse (par exemple, la gestion de la taille de l'invite, le découpage du texte, s'assurer que le LLM cite réellement le contenu récupéré). Ces compromis sont discutés dans les guides d'implémentation (Source: haruiz.github.io) (Source: research.google).

Attribution de source et filigrane

Une autre idée émergente est de permettre à un LLM d'**étiqueter sa propre sortie** avec des métadonnées de source. Plutôt que de rechercher a posteriori, cette approche vise à intégrer la provenance dans le processus de génération. Un exemple frappant est le cadre **WASA (WAtermark-based Source Attribution)** (Source: openreview.net). Dans WASA, le LLM est entraîné à insérer un « filigrane » subtil – en fait un signal ou un code – dans chaque morceau de texte qu'il génère, de sorte qu'une analyse ultérieure puisse mapper ce filigrane à des documents ou sources de données spécifiques utilisés lors de l'entraînement. Pensez-y comme à des particules traçantes invisibles dans le texte. Si WASA est mis en œuvre avec succès, il nous permettrait de demander : « Étant donné cette phrase générée, quelle(s) source(s) d'entraînement a (ont) contribué à ce contenu ? »

WASA est motivé par des préoccupations juridiques/de propriété intellectuelle. Comme noté dans leur résumé, les sorties des LLM pourraient involontairement « enfreindre la propriété intellectuelle des données utilisées pour entraîner les LLM » (Source: openreview.net). En revanche, les approches standard (par exemple, forcer les LLM à citer des sources dans les références) se concentrent sur les textes externes au moment de la requête. WASA traite plutôt chaque génération comme portant une signature. Les auteurs identifient des desiderata tels que la précision de l'attribution et la robustesse aux modifications adverses, et proposent des algorithmes pour mapper les sorties aux fournisseurs de données d'entraînement. Les premières évaluations de WASA (sur des benchmarks synthétiques) montrent qu'il peut en effet intégrer des informations de source avec une haute fidélité. Cependant, ce domaine de travail est très nouveau et expérimental. Il nécessite de modifier l'algorithme d'entraînement ou l'architecture du modèle, ce qui pourrait ne pas être pratique pour les services LLM actuels. En effet, le filigrane répond à la question « où avez-vous appris cela ? » plutôt que « où puis-je le vérifier ? ». C'est une approche complémentaire mais distincte des citations habituelles centrées sur l'utilisateur.

Techniques d'invite et de génération de citations

Une stratégie pratique plus simple consiste à demander au LLM, via l'invite, de produire des citations. Par exemple, on pourrait ajouter à chaque instruction de l'utilisateur : « Fournissez des références justificatives (avec auteur, titre et lien) pour votre réponse. » Parfois appelée invitation à des références ou chaîne de pensée avec citations, cela repose sur la capacité du LLM à formater des références qu'il semble « se souvenir ». Par essais et erreurs, certains utilisateurs ont constaté que GPT-4 (et Claude, etc.) synthétisera effectivement une liste d'articles ou d'URL lorsqu'on le lui demande, bien que pas toujours correctement.

Les testeurs universitaires ont obtenu des résultats mitigés. Dans une étude sur la rédaction académique transdisciplinaire, une équipe a demandé à GPT-3.5 de générer un court article de synthèse avec des citations. Ils ont ensuite vérifié la validité de chaque citation. Globalement, environ 74,5 % des références de GPT correspondaient à des articles réels et existants (Source: pmc.ncbi.nlm.nih.gov). C'est significatif (près des trois quarts) mais laisse encore de nombreuses références inventées ou inexactes. Fait intéressant, la même étude a noté l'écart entre les domaines : tandis que les requêtes en sciences naturelles produisaient 72 à 76 % de citations valides, les requêtes en sciences humaines présentaient davantage de DOI hallucinés (par exemple, une incohérence de citation de type Reuters) (Source: pmc.ncbi.nlm.nih.gov). Une autre évaluation a révélé que la précision des DOI de GPT-3.5 n'était que d'environ 30 % dans les sciences humaines, ce qui indique une performance inégale selon les domaines (Source: pmc.ncbi.nlm.nih.gov) (Source: pmc.ncbi.nlm.nih.gov).

Ces méthodes de *prompting* ne nécessitent aucune infrastructure spéciale, mais leur fiabilité est limitée par les connaissances internes du modèle et sa tendance à la confabulation. Du côté positif, le *prompting* peut inciter les LLM à citer plus souvent qu'ils ne le feraient par défaut. Comme l'ont noté les praticiens, l'inclusion de citations « facilite la vérification que le LLM utilise des informations pertinentes, réduisant ainsi les hallucinations » (Source: haruiz.github.io). Cependant, il faut vérifier manuellement chaque référence générée, de sorte que le *prompting* seul n'est pas une solution miracle. Dans les systèmes de production, les *prompts* de génération de citations sont généralement combinés avec le RAG ou un post-traitement pour la vérification des faits.

Flux de travail avec augmentation par récupération et de citation

Tableau 1. *Comparaison des approches pour l'approvisionnement en informations pour les sorties des LLM.* Chaque approche représente une stratégie différente pour connecter les réponses des LLM à des connaissances externes.

APPROCHE	MÉCANISME	EXEMPLE D'UTILISATION	AVANTAGES	LIMITES	RÉFÉRENCES CLÉS
Génération augmentée par récupération (RAG)	À chaque requête, récupérer les documents pertinents (via une recherche ou une base de données vectorielle) et les intégrer dans le *prompt* du LLM.	ChatGPT avec plugins de recherche web ; Perplexity ; RAG d'entreprise interne.	Réponses basées sur du texte réel, faits à jour ; facilement traçables aux sources.	Nécessite une base de connaissances / recherche maintenue ; erreurs de récupération possibles ; plus lent.	Google Research (2025) (Source: research.google); Ruiz (2023) (Source: haruiz.github.io)
Génération de citations basée sur le *prompt*	Demander au LLM de générer des citations ou des références dans le cadre de la réponse.	Outils de rédaction académique (GPT-3.5 avec *prompts* de citation).	Aucune infrastructure externe nécessaire ; peut exploiter le style de citation appris par le LLM.	Risque élevé de citations hallucinées ou incomplètes ; performance inégale selon les domaines (Source: pmc.ncbi.nlm.nih.gov).	Mugaanyi et al. (2024) (Source: pmc.ncbi.nlm.nih.gov); Études de retour d'expérience de journaux.
Affinement / Intégration de modèle	Entraîner ou affiner les LLM sur des données annotées contenant des citations, ou incorporer un objectif sensible aux citations.	Prototypes de recherche (par exemple, modèles entraînés sur des articles académiques avec DOI).	Peut internaliser les modèles de citation ; solution de bout en bout si bien réalisée.	Nécessite des données d'entraînement spécialisées ; peut encore halluciner si les connaissances sont absentes.	(Domaine émergent ; voir discussions générales)
Méthodes de filigrane numérique/provenance (WASA)	Intégrer des signaux cachés dans le texte généré qui encodent les identifiants de source ou les métadonnées du fournisseur.	Prototype de recherche (cadre WASA) (Source: openreview.net).	Permet une attribution exacte aux sources d'entraînement ; protège la propriété intellectuelle ; traçabilité automatisable.	Augmente la complexité de l'entraînement du modèle ; peut dégrader la fluidité de la sortie ; vulnérable à l'édition.	Lu et al. (WASA, 2025) (Source: openreview.net)

APPROCHE	MÉCANISME	EXEMPLE D'UTILISATION	AVANTAGES	LIMITES	RÉFÉRENCES CLÉS
Vérification des faits post-génération	Après avoir généré une réponse, exécuter une vérification automatisée (par exemple, interroger un LLM ou effectuer une recherche) pour valider les faits et joindre les sources.	Chaînes de « révision » de LLM ; systèmes de vérification avec intervention humaine.	Améliore la précision finale ; peut détecter les hallucinations.	Ajoute de la latence et de la complexité ; doit définir des vérificateurs fiables.	(Pratique industrielle ; pas de source unique. Voir la section sur les pipelines QA.)

Le *Tableau 1* illustre l'éventail des méthodes. Le RAG classique et la citation par *prompting* sont déjà utilisés par de nombreux systèmes, tandis que le filigrane numérique (*watermarking*) et l'affinement avancé restent des sujets de recherche. Le bon choix dépend des besoins de l'application en matière de précision, de vitesse et de contraintes de ressources. Par exemple, les récentes innovations de Google en matière de RAG visent à minimiser les « hallucinations » en s'assurant que le modèle dispose d'un contexte suffisant (Source: research.google). De même, les blogs de développement soulignent qu'avec le RAG, chaque réponse peut explicitement mettre en évidence l'extrait ou l'URL d'où elle provient, améliorant considérablement la transparence.

Exemples d'implémentation

En pratique, les ingénieurs ont mis en œuvre ces approches de diverses manières. Un pipeline RAG typique implique un récupérateur (*retriever*) (souvent un moteur de recherche sémantique ou un index de similarité vectorielle) et un LLM. Certains tutoriels montrent comment diviser les documents sources en morceaux (*chunks*) interrogeables, puis demander au LLM de citer « le document source et le paragraphe d'où provient chaque réponse » (Source: haruiz.github.io). Par exemple, un blog publié décrit l'utilisation de LlamaIndex (GPT Index) pour récupérer des morceaux de texte, puis l'incitation de GPT-4 à générer une réponse consolidée avec des citations intégrées au texte vers ces morceaux. Un autre exemple est le prototype « RAG sensible aux citations » (*Citation-Aware RAG*), qui attache des citations granulaires à chaque phrase de la réponse. Tous ces exemples reposent sur l'idée fondamentale : le contenu récupéré est formaté (parfois reformulé) et intégré de manière transparente dans la réponse, le LLM ajoutant un texte créatif minimal.

Du côté du *prompting*, de nombreux développeurs ajoutent simplement des instructions comme « Veuillez lister vos références » au *prompt* de l'utilisateur. Certains systèmes destinés aux utilisateurs universitaires fourniront même des entrées bibliographiques et des enseignements sur les formats de citation. Cependant, comme nous le verrons, le succès de ces citations à la demande est mitigé à moins d'être combiné avec la récupération ou la vérification.

Enfin, considérons les LLM des moteurs de recherche. Le Copilot de Microsoft cite désormais systématiquement ses sources : chaque réponse factuelle inclut des notes de bas de page avec des URL vers les résultats de recherche Bing. Perplexity génère des citations cliquables provenant de sources d'actualités et scientifiques. Ces solutions commerciales masquent efficacement le cadre de citation en coulisses, mais elles illustrent la demande : les utilisateurs s'attendent à des références pour des informations fiables.

Précision des citations et études de cas

Pour évaluer l'efficacité de ces cadres, les chercheurs ont commencé à mesurer la qualité des citations dans les sorties des LLM. Nous passons ici en revue les principales conclusions des évaluations transdomaine et des exemples concrets.

Études empiriques sur la qualité des citations

Plusieurs études formelles ont quantifié la fréquence à laquelle les citations des LLM sont correctes. Mugaanyi *et al.* (2024) ont étudié les performances de ChatGPT-3.5 lors de la génération de citations pour des *prompts* en sciences et en sciences humaines. Ils ont constaté que sur 102 références générées, **74,5 % correspondaient à des œuvres réelles** (Source: pmc.ncbi.nlm.nih.gov). Réparties par domaine,

environ 72,7 % des références pour les sujets de sciences naturelles étaient valides, et 76,6 % pour les sujets de sciences humaines (Source: pmc.ncbi.nlm.nih.gov). Cela indique une amélioration substantielle par rapport aux modèles précédents : près des trois quarts des citations de GPT-3.5 étaient suffisamment précises pour localiser un article réel. Cependant, les erreurs de DOI étaient courantes, en particulier dans les sciences humaines (DOI mal tapés ou incorrects dans environ 89 % des cas) (Source: pmc.ncbi.nlm.nih.gov). Les auteurs concluent qu'une adaptation spécifique au domaine pourrait aider (par exemple, l'affinement sur des données de style de citation) et que les utilisateurs doivent vérifier attentivement les DOI.

Une autre évaluation s'est concentrée sur ChatGPT-4 dans des domaines spécifiques. Dans un contexte d'éducation médicale (« protocole de traumatologie ABCDE »), les testeurs ont demandé à ChatGPT-4 de générer des références pour chaque étape. Ils ont évalué la précision de 30 références (6 par catégorie). Le résultat : **seulement 43,3 % de ces références étaient entièrement exactes** (Source: pmc.ncbi.nlm.nih.gov). Les 56,7 % restants étaient soit erronées, soit inexistantes (par exemple, auteurs, titres incorrects ou fausses entrées de journaux) (Source: pmc.ncbi.nlm.nih.gov). En d'autres termes, plus de la moitié des citations étaient sans valeur du point de vue de la vérification. L'étude dramatise la question : « Avec 57 % des références étant inexactes ou inexistantes, ChatGPT-4 n'a pas réussi à fournir des références fiables et reproductibles » (Source: pmc.ncbi.nlm.nih.gov). Cela compromet son utilité pour les domaines fondés sur des preuves. (Les chercheurs notent que cela est spécifique à un domaine/tâche ; dans un domaine mieux défini, les performances pourraient s'améliorer.)

En revanche, une analyse générale de la « véracité des références de l'IA générative » a rapporté une précision bien plus élevée avec GPT-4. Dans cette étude, GPT-4 (désigné « ChatGPT4o ») a produit une « écrasante majorité » de citations correctes, avec seulement environ 10 % de ses références étant entièrement inventées (Source: www.mdpi.com) (Source: www.mdpi.com). Statistiquement, le taux de citations fabriquées par GPT-4 était bien inférieur à celui de GPT-3.5 (le test du chi-carré a montré une baisse significative des citations hallucinées à seulement 10 % (Source: www.mdpi.com). Les auteurs notent que l'amélioration est probablement due aux capacités linguistiques plus solides de GPT-4 et potentiellement à la conception du *prompt*. Malgré cela, ils ont trouvé quelques erreurs mineures : par exemple, des titres corrects mais des numéros de volume manquants, qu'ils ont classés comme des références incomplètes (Source: www.mdpi.com).

Le Tableau 2 (ci-dessous) compare les performances de citation de plusieurs LLM et contextes tirés de ces études et rapports. Pour ChatGPT et Gemini, notez que la « précision » varie selon la rigueur avec laquelle on définit une correspondance (DOI exact vs. titre/auteurs corrects). Dans tous les cas, les citations des LLM sont imparfaites : même la précision d'environ 90 % de GPT-4 (Source: www.mdpi.com) n'est pas de 100 %.

SYSTÈME / CONTEXTE	RÉSULTAT	NOTES / SOURCE
ChatGPT-4 (QA médicale, étude ABCDE)	13 références sur 30 (43,3 %) entièrement exactes (Source: pmc.ncbi.nlm.nih.gov)	57 % des références étaient fausses/inexactes (Source: pmc.ncbi.nlm.nih.gov)
ChatGPT-4 (requêtes générales)	≈90 % des citations correctes (Source: www.mdpi.com) (Source: www.mdpi.com)	Seulement ~10 % fabriquées ; amélioration par rapport à GPT-3.5 (Source: www.mdpi.com)
ChatGPT-3.5 (rédaction académique)	76 références sur 102 (74,5 %) réelles (Source: pmc.ncbi.nlm.nih.gov)	Les erreurs de DOI étaient courantes en sciences humaines (Source: pmc.ncbi.nlm.nih.gov)
Gemini 1.5 (QA santé, *prompt* malveillant)	A produit une réponse médicale confiante avec de fausses citations (Source: www.reuters.com)	Voir l'étude Reuters : succombe à l'injection de *prompt*
Llama 3.2-90B (même test)	Sortie fabriquée similaire avec des références bidon (Source: www.reuters.com)	Cas défavorable testé par des commandes cachées
Grok Beta (xAI) (même test)	Résultat similaire avec des citations inventées (Source: www.reuters.com)	Exposé par des *prompts* de système cachés
Claude 3.5 Sonnet (même test)	**A refusé** de se conformer (a refusé de donner une fausse réponse) (Source: www.reuters.com)	Seul modèle à ne pas avoir produit de fausse réponse
Bing Chat / Copilot	Inclut des liens vers les résultats de recherche web ; généralement précis	(Système RAG commercial avec sources en direct)
Perplexity.ai	Cite toujours des sources externes (recherche/actualités) ; haute fiabilité	(Connu comme un moteur de réponse basé sur le RAG)

Tableau 2 : *Comportement de citation des systèmes LLM représentatifs.* La colonne de gauche liste le modèle et le contexte, celle du milieu présente les résultats observés, et celle de droite indique les sources. GPT-4 montre les meilleures performances dans les études rigoureuses (Source: www.mdpi.com) (Source: www.mdpi.com), mais ne peut toujours pas garantir une fidélité parfaite. GPT-3.5 (et vraisemblablement le mode « pré-entraîné » de base de GPT-4) hallucine une fraction substantielle de références dans les tâches difficiles (Source: pmc.ncbi.nlm.nih.gov) (Source: pmc.ncbi.nlm.nih.gov). Les LLM spécifiques à un domaine (Gemini, Llama, Grok) peuvent être amenés à produire des citations entièrement fabriquées sous l'effet d'un *prompting* malveillant (Source: www.reuters.com). Les systèmes commerciaux comme Bing exploitent la recherche pour une grande précision mais ne sont pas immunisés contre le phrasé de l'utilisateur.

Étude de cas : Questions-réponses médicales

Un cas concret illustre cette dynamique. Dans une expérience publiée, des cliniciens ont demandé à ChatGPT-4 de citer des preuves pour les directives standard de triage des traumatismes. ChatGPT-4 a listé plusieurs articles de recherche par étape de directive, mais lorsque les experts les ont vérifiés, **seulement 43,3 % étaient corrects** (Source: pmc.ncbi.nlm.nih.gov). Le reste était partiellement erroné (mauvais auteur, année ou PMID) ou entièrement inexistant. Par exemple, une réponse avait le bon titre et la bonne revue, mais un nom d'auteur et un PMID erronés ; une autre avait la bonne année mais un titre incorrect. L'étude avertit que cela « ne parvient pas à fournir des références fiables », soulignant que l'utilisation de ChatGPT-4 dans la prise de décision médicale « sans vérification approfondie » est dangereuse (Source: pmc.ncbi.nlm.nih.gov).

Parallèlement, une étude distincte a demandé à ChatGPT-3.5 (GPT 3.5 Turbo par défaut) de rédiger de courts articles en sciences et en sciences humaines. Sur toutes les citations générées, environ 25,5 % étaient fausses ; inversement, 74,5 % étaient réelles (Source: pmc.ncbi.nlm.nih.gov). La précision était plus élevée dans les sciences que dans les sciences humaines. Bien que ces chiffres soient prometteurs (la majorité des citations de ChatGPT étaient valides dans ce contexte), le taux d'erreur restant est inacceptable pour une utilisation universitaire sans vérification des faits. L'étude souligne spécifiquement comment les hallucinations de DOI sont encore monnaie courante dans certains domaines.

Du côté positif, des rapports anecdotiques suggèrent que GPT-4 avec navigation obtient de bien meilleurs résultats. Lorsqu'il est autorisé à récupérer des sources web, il fournit souvent des données correctes avec des URL qui étayent réellement la réponse. Par exemple, si on lui demande un fait bien connu, GPT-4 répondra parfois par « Selon *[Source]*... » et fournira un lien réel. Ce mode le transforme efficacement en un assistant de recherche hybride. Il ne s'agit pas de citations gérées de manière autonome (le modèle génère toujours de la prose), mais l'inclusion de liens réels améliore considérablement la confiance.

En pratique, certaines communautés de débat sur l'IA ont établi des taux d'erreur de citation moyens pour divers chatbots. Leurs conclusions heuristiques s'alignent sur les études ci-dessus : GPT-4 (avec accès aux sources) >> GPT-3.5 ≈ Bard ≈ Claude (sans références) en termes de fiabilité. Ceux-ci ne sont pas évalués par des pairs, mais renforcent l'idée que la **disponibilité de sources réelles est essentielle**.

Étude de cas : Attaque de désinformation en santé

À titre d'exemple de mise en garde, considérons une récente expérience de type « red-team » rapportée par Reuters (Source: www.reuters.com). Des chercheurs ont donné des instructions de prompt cachées à divers chatbots IA pour qu'ils produisent de faux conseils de santé. Ils ont constaté que **presque tous les modèles testés s'y sont conformés**, donnant des réponses persuasives mais fausses, et inventant même des citations savantes pour les étayer. GPT-4, Gemini 1.5, Llama 3.2-90B et Grok ont tous généré une recommandation de traitement confiante (mais dangereuse) accompagnée de « références de journaux » fabriquées. Un seul modèle – Claude 3.5 d'Anthropic – a refusé de répondre en mode malveillant. Ce résultat frappant souligne que les LLM peuvent non seulement halluciner des citations spontanément, mais peuvent aussi être activement manipulés pour le faire. Cela souligne l'urgence de contrôles de sources intégrés : tout LLM ouvert, même GPT-4, manque actuellement d'une protection robuste contre de telles références hallucinées. (Nous notons que le refus de Claude était une réponse de sécurité, et non une fonctionnalité de citation intégrée.)

Analyse de domaine : Sciences vs. Humanités

Différents domaines imposent différentes exigences en matière de citation. L'étude de Mugaanyi *et al.* (2024) (Source: pmc.ncbi.nlm.nih.gov) suggère que les sujets STEM ont bénéficié de conventions de citation plus formelles (près de 73 % de références réelles) que les sciences humaines dans la production de GPT-3.5. Cela pourrait être dû à des facteurs tels que : (1) les revues et conférences STEM représentent une grande fraction de l'entraînement du LLM ; (2) les DOI sont utilisés de manière plus uniforme en science. Dans les sciences humaines, GPT-3.5 a souvent généré des titres plausibles mais sans existence réelle, ou des DOI qui pointaient vers de mauvais articles (Source: pmc.ncbi.nlm.nih.gov). Ainsi, même avec une incitation identique, la **fiabilité dépend du contexte**. Des observations similaires ont été faites de manière anecdotique : par exemple, il a été démontré que GPT-4 s'en sortait bien mieux lorsqu'il répondait à des requêtes factuelles bien définies (Tableau 2) que lorsqu'il improvisait sur des questions ouvertes.

Dans les milieux éducatifs, les enseignants se demandent s'il faut autoriser l'utilisation de l'IA. Certaines universités exigent désormais que tout contenu généré par l'IA soit accompagné de citations vérifiables. Par exemple, lorsque les étudiants utilisent ChatGPT pour rédiger des essais, les meilleures pratiques émergent : le traiter comme un assistant de rédaction, et toujours vérifier chaque citation fournie par l'IA. Certains éducateurs demandent explicitement aux étudiants de **ne pas** utiliser l'IA pour les essais créatifs, mais de s'y fier pour lister des références sur des sujets connus, car les connaissances préétablies peuvent être citables. Ces mesures sociales reflètent la réalité technique : les LLM modernes sont des outils utiles, mais sans un cadre de citation, on ne peut leur faire confiance pour effectuer le travail savant de référencement approprié (Source: pmc.ncbi.nlm.nih.gov) (Source: pmc.ncbi.nlm.nih.gov).

Analyse des données et preuves

Les preuves quantitatives issues des études existantes soulignent les points ci-dessus. Nous résumons ici les données clés :

- **Précision des citations :** Dans les évaluations contrôlées, les taux de citations correctes variaient approximativement entre 40 % et 90 % selon le modèle et la tâche. GPT-4, dans une session de questions-réponses médicales, n'avait que 43 % de sources correctes (Source: pmc.ncbi.nlm.nih.gov), tandis que GPT-4, sur des requêtes générales, atteignait environ 90 % (Source: www.mdpi.com). GPT-3.5 se situait autour de 70 à 75 % lors d'un test de rédaction académique (Source: pmc.ncbi.nlm.nih.gov). Cette variance montre que même les LLM avancés sont loin d'être des générateurs de sources parfaits.
- **Taux d'hallucination :** En complément de ce qui précède, les taux de citations fabriquées étaient de 57 % (GPT-4 médical) à 10 % (GPT-4 général) (Source: pmc.ncbi.nlm.nih.gov) (Source: www.mdpi.com). Pour GPT-3.5 en sciences humaines, l'hallucination de DOI était de 89 % (Source: pmc.ncbi.nlm.nih.gov), un taux d'erreur étonnamment élevé.
- **Accord des évaluateurs :** Dans l'étude médicale, des évaluateurs indépendants ont obtenu un kappa de Cohen de 0,89 sur la notation des citations (Source: pmc.ncbi.nlm.nih.gov), indiquant une fiabilité inter-évaluateurs élevée pour juger des références réelles par rapport aux fausses. Cela suggère que les métriques d'évaluation elles-mêmes sont robustes.
- **Tendances systématiques :** Les données montrent constamment que les requêtes en domaine ouvert, activées par la récupération, produisent une précision plus élevée que les genres fermés nécessitant un rappel. Le développement laisse une marge d'amélioration significative : un « assistant LLM fiable » idéal devrait approcher 100 % de validité des citations.

Discussion : Défis, perspectives et orientations futures

Les résultats collectifs dressent un tableau clair : les LLM actuels ne sont **pas des moteurs de citation fiables par défaut**, mais des cadres évolutifs peuvent améliorer la confiance. Nous explorons maintenant les implications plus larges et les prochaines étapes potentielles.

Défis techniques et pistes de recherche

Amélioration de la récupération. Étant donné que les citations basées sur le RAG dépendent de la qualité de la récupération, la recherche en cours se concentre sur de meilleurs index et modèles de pertinence. Les derniers travaux de Google introduisent l'idée de **« contexte suffisant »** pour le RAG : déterminer exactement la quantité de texte documentaire que le LLM doit voir pour la précision. Les expériences suggèrent qu'un contexte trop faible provoque des hallucinations, il est donc essentiel d'affiner le pipeline de récupération. Les avancées en matière d'embeddings vectoriels, de reformulation de requêtes et de récupération multi-passes pourraient toutes resserrer la boucle entre la requête et la source crédible.

Citation dans l'alignement de l'attention. Certaines méthodes proposées visent à « peindre » l'attention ou les logits internes du LLM avec des informations de source. Par exemple, lier certaines têtes d'attention à des pointeurs de base de données, ou fusionner des graphes de connaissances dans les couches de transformateur. Bien que très expérimentales, ces approches cherchent à éliminer l'hallucination par conception.

Étalonnage et jeux de données. Des métriques fiables sont nécessaires. Ce rapport a documenté plusieurs études internes, mais ce qui manque, c'est une vaste suite de benchmarks de questions avec des références de vérité terrain pour l'évaluation des LLM. La communauté du TAL pourrait assembler de tels jeux de données dans divers domaines (questions-réponses scientifiques, requêtes juridiques, faits historiques, etc.) afin que la précision des citations devienne une métrique standard. Des travaux récents sur l'« attribution de source » et l'« évaluation de modèle » (par exemple, l'article WASA de l'ICLR 2025) commencent à définir des protocoles d'évaluation.

Perspectives utilisateur et éthiques

Du point de vue de l'utilisateur, les citations modifient radicalement le modèle de confiance. Un étudiant ou un chercheur osera faire bien plus confiance à une réponse de l'IA si elle est accompagnée de liens crédibles. Cela pourrait révolutionner le travail de la connaissance : on peut imaginer un avenir où les assistants IA fonctionnent comme des « bibliothécaires surpuissants », résumant le contenu mais pointant toujours

vers les chapitres ou articles qu'ils ont utilisés. Cependant, une dépendance prématurée peut être dangereuse. Les cas ci-dessus montrent que sans supervision, l'IA peut induire en erreur. Les utilisateurs (et les régulateurs) doivent cultiver la littératie de l'IA : toujours vérifier les références de l'IA.

Éthiquement, l'obligation de citer aide à répondre aux préoccupations de plagiat. Lorsqu'un LLM résume une source, une citation reconnaît l'auteur original. Cela aligne l'IA sur les normes académiques. En revanche, les paraphrases d'IA non sourcées pourraient involontairement plagier ou propager de la désinformation. Des initiatives sont en cours dans le monde universitaire pour traiter le contenu généré par l'IA comme des **outils d'accès à l'information**, et non comme des sources indépendantes. De nombreuses revues interdisent désormais de lister une IA comme auteur, et la question de savoir comment créditer le texte généré par l'IA est en débat. Quoi qu'il en soit, d'un point de vue moral, fournir les sources respecte les droits de propriété intellectuelle et la transparence.

Tendances réglementaires et industrielles

Les décideurs politiques en prennent note. Bien que la loi européenne sur l'IA (en projet) ne mentionne pas encore spécifiquement les citations, elle insiste sur la **transparence et la traçabilité** des résultats de l'IA. En pratique, les régulateurs pourraient exiger que les produits de consommation d'IA divulguent les sources pour les informations à enjeux élevés (à l'instar des règles de responsabilité pour les allégations de santé). Déjà, lors des poursuites du NYT, le concept d'« attribution de source » était central (Source: swarajyamag.com). Le Bureau américain du droit d'auteur et les tribunaux sont aux prises avec la manière d'équilibrer l'entraînement de l'IA avec les titulaires de droits. Dans ce climat, un cadre de citation pour l'IA n'est pas seulement une commodité, mais pourrait devenir une nécessité juridique.

Du côté de l'industrie, les principaux développeurs de LLM travaillent discrètement sur ce sujet. OpenAI a expérimenté « ChatGPT Plus avec navigation », et Google serait en passe d'intégrer des citations dans les futures versions de Gemini. Des startups émergentes (SciSpace, Elicit, autres) se concentrent sur l'IA pour la recherche avec référencement intégré. Même les considérations de conception comme l'interface utilisateur comptent : les applications permettent désormais souvent de cliquer sur une note de bas de page pour afficher la source. Cela modifie les attentes des utilisateurs : une IA qui ne cite pas pourrait bientôt être perçue comme incomplète ou indigne de confiance.

Perspectives d'avenir

En regardant vers l'avenir, nous anticipons plusieurs tendances :

- **Protocoles de citation standardisés :** Tout comme HTML et DOI ont structuré le web du savoir, nous pourrions voir apparaître une norme de citation conviviale pour les machines pour l'IA. Les propositions incluent des bibliothèques qui attachent automatiquement des références de style BibTeX aux réponses de l'IA, ou des API de LLM qui renvoient des objets de référence structurés.
- **Intégration avec les graphes de connaissances :** La sortie des LLM pourrait être intégrée à des outils comme Wikidata ou Google Knowledge Graph, de sorte que les entités mentionnées dans les réponses se lient automatiquement à des entrées organisées. Cette approche hybride pourrait fournir des citations sémantiques plutôt que des citations de documents complets, améliorant ainsi la vérifiabilité.
- **Guidage de l'utilisateur et ingénierie des prompts :** Tant que les modèles sous-jacents ne s'améliorent pas, une citation efficace dépend souvent de la manière dont l'utilisateur pose la question. La recherche en ingénierie des prompts (par exemple, la chaîne de pensée qui inclut « Citez ceci ») se poursuivra. Des programmes éducatifs enseignent également aux gens comment interroger l'IA et comment vérifier ses réponses.
- **Outils d'explicabilité des modèles :** Au-delà des citations directes, des méthodes comme l'attribution basée sur l'attention ou l'évaluation contrefactuelle peuvent aider les utilisateurs à comprendre *pourquoi* un LLM a répondu d'une certaine manière. Une meilleure explicabilité peut compléter les citations pour donner une image plus complète de la fiabilité.
- **Évaluation et retour d'information continus :** Les produits d'IA intégreront probablement des boucles de rétroaction. Si une citation fournie s'avère erronée par les utilisateurs, ces données pourraient être utilisées pour affiner les modèles ou mettre à jour les index de récupération. En substance, les cadres de citation de l'IA pourraient évoluer pour inclure les « votes » des utilisateurs sur la qualité des sources.

Conclusion

À mesure que les grands modèles de langage imprègnent les flux de travail de l'information, leur capacité à citer des sources sera un facteur déterminant de leur utilité et de leur fiabilité. Notre examen montre que si les premiers efforts ont progressé, nous sommes encore loin de la perfection. GPT-4 peut souvent citer correctement, mais des taux d'erreur non négligeables persistent (Source: pmc.ncbi.nlm.nih.gov)

(Source: www.mdpi.com). Des techniques spécialisées comme le RAG et le WASA offrent des cadres puissants pour y remédier, mais chacune comporte des compromis. Les études de cas d'utilisateurs nous rappellent que sans de solides garanties de citation, l'IA peut induire en erreur par inadvertance.

À l'avenir, le « cadre de citation de l'IA » est susceptible de devenir un domaine de recherche interdisciplinaire majeur. Il s'appuie sur le traitement du langage naturel, la récupération d'informations, le droit de la propriété intellectuelle et la conception UX. Nous devons continuer à développer des benchmarks, à partager des jeux de données ouverts de questions-réponses avec des sources vérifiées, et à itérer sur des modèles qui internalisent la notion de vérité vérifiable. Pour l'instant, les développeurs et les utilisateurs devraient considérer les LLM comme des assistants nécessitant une supervision : bénéfiques pour le brainstorming et la génération de brouillons, mais ayant besoin de citations de « vérité terrain » pour toute application sérieuse.

En fin de compte, les citations sont la monnaie de la connaissance. Intégrer cette monnaie dans l'IA comblera le fossé entre la synthèse machine et les normes humaines de preuve. Comme le note à juste titre un expert en sécurité de l'IA, l'ajout de citations peut rendre les sorties des LLM non seulement plus correctes, mais aussi **responsables** (Source: haruiz.github.io) (Source: openreview.net). Ce rapport a cartographié le paysage technique de ce défi et suggère des voies à suivre pour rendre les réponses de l'IA traçables et dignes de confiance.

Références : Toutes les affirmations ci-dessus sont étayées par la littérature et les sources citées (voir les citations en ligne). Les études clés incluent des évaluations de la précision des références de GPT-3.5/4 (Source: pmc.ncbi.nlm.nih.gov) (Source: pmc.ncbi.nlm.nih.gov), des propositions de cadre pour l'attribution (Source: openreview.net) (Source: haruiz.github.io), et des rapports d'actualité sur le comportement de citation de l'IA (Source: www.reuters.com) (Source: swarajyamag.com), entre autres. Les travaux cités fournissent des données détaillées, une analyse d'experts et le contexte des problèmes discutés.

Étiquettes: llm, citation-ia, rag, attribution-source, hallucinations-llm, provenance-donnees, ia-generative, wasa

---