De la génération à la désinformation : une étude critique des hallucinations dans les LLMs

Katia Lounas, Edgar Demeude

Mai 2025

Resumé

Les modèles de langage de grande taille (LLMs) ont transformé la génération automatique de texte, atteignant des niveaux de fluidité et de cohérence sans précédent. Toutefois, leur capacité à produire des textes factuellement erronés ou trompeurs, communément désignée sous le terme de hallucination, soulève des préoccupations en matière de fiabilité et constitue un obstacle majeur à leur adoption dans des domaines sensibles comme la médecine, le droit ou l'éducation. Ce phénomène soulève une question fondamentale : ces modèles peuvent-ils véritablement être considérés comme des générateurs fiables de savoir ? Dans cet article, nous proposons une revue critique des causes sous-jacentes des hallucinations, qu'elles proviennent des limites des données d'entraînement, de l'architecture des modèles, ou des méthodes de génération utilisées. Nous présentons les principales typologies des hallucinations identifiées dans la littérature, ainsi que les approches récentes visant à les atténuer, notamment via l'intégration de systèmes externes de mémoire ou de récupération d'information (retrieval-augmented generation). Enfin, nous discutons des limites actuelles des protocoles d'évaluation et soulignons les défis méthodologiques que pose la mesure fine de la factualité. Cet article s'attache à mettre en lumière les limites actuelles de la fiabilité des modèles de langage et à nourrir la réflexion autour du développement de systèmes de génération textuelle plus rigoureux, transparents et dignes de confiance.

Mots-clés : modèles de langage, hallucination, intelligence artificielle, LLMs, évaluation, biais.

1 Introduction

Les LLMs ont connu une adoption massive dans de nombreux secteurs, grâce à leur capacité à générer des textes cohérents et créatifs. Cependant, malgré leurs performances impressionnantes, ces modèles présentent des limitations intrinsèques qui soulèvent des questions sur leur fiabilité et leur utilisation responsable, les hallucinations. Ce phénomène correspond à la production, par le modèle, de contenus incorrects, infondés ou inventés, présentés néanmoins comme fiables. Les LLMs, en tant que prédicteurs statistiques, ne possèdent ni compréhension du réel, ni mécanisme natif de vérification des faits. Ils peuvent ainsi propager des informations erronées de manière confiante, ce qui pose des risques importants, notamment dans des domaines sensibles comme la santé, le droit ou l'éducation. Dès lors, il devient essentiel de comprendre non seulement ce qu'est une hallucination, mais aussi pourquoi elle se produit et comment on peut la détecter ou la limiter.

2 Qu'est-ce qu'une hallucination dans un LLM ?

Le terme hallucination est couramment utilisé dans le domaine du traitement automatique du langage naturel (TALN) pour désigner la génération de contenu textuel qui est soit incohé-

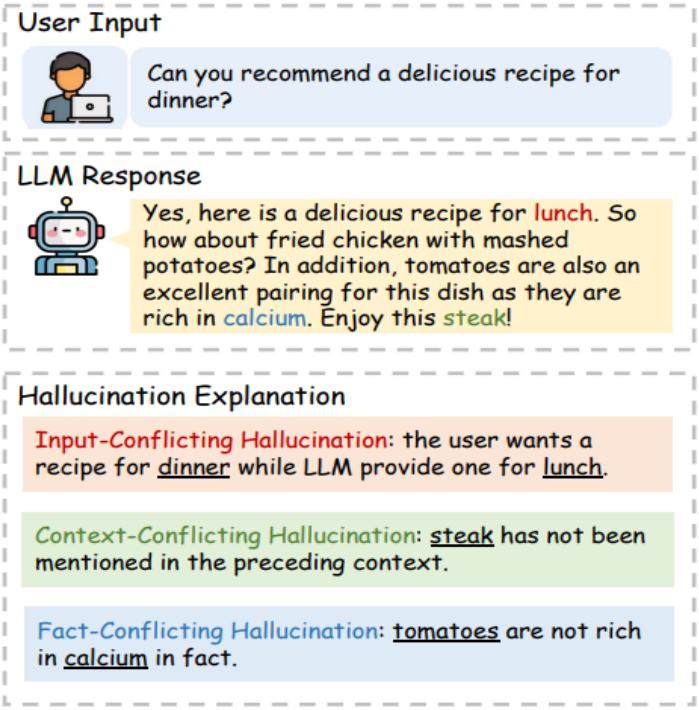

FIGURE 1 – Trois types d'hallucinations apparaissent dans les réponses des LLMs[17]

rent, soit déconnecté de la source d'information d'origine [7]. On distingue généralement deux types d'hallucinations : les hallucinations intrinsèques et les hallucinations extrinsèques. Nous reviendrons sur les différentes typologies dans la section suivante.

Zhang et al. [17] ont proposé une typologie plus fine adaptée aux LLMs modernes, distinguant trois types de conflits (voir FIG 1) : (1) conflit avec l'entrée (input-conflicting), lorsque le texte s'éloigne de l'instruction ou de la donnée initiale ; (2) conflit avec le contexte (context-conflicting), où les réponses contredisent des éléments générés précédemment ; et (3) conflit avec le savoir (fact-conflicting), lorsque le modèle produit des informations erronées par rapport aux connaissances établies.

3 État de l'art

La problématique des hallucinations génératives a suscité un intérêt croissant dans les travaux récents sur les LLMs. Plusieurs études ont tenté de formaliser cette problématique et d'en proposer des typologies. Zhang et al. (2023) [17], par exemple, affinent la distinction classique entre hallucinations intrinsèques et extrinsèques en identifiant des sous-catégories spécifiques aux LLMs, telles que les conflits avec l'entrée, le contexte généré ou les connaissances factuelles (voir FIG 1).

D'autres travaux, comme celui de Liu et al. (2023) [8], se concentrent sur l'évaluation comparative des modèles, mettant en lumière le paradoxe selon lequel des systèmes linguistiquement plus performants ne sont pas nécessairement plus fiables sur le plan factuel. Ces analyses soulignent l'importance d'outils de mesure normalisés pour évaluer la véracité des sorties générées.

Face à ces limites, plusieurs pistes sont explorées pour réduire les hallucinations, notamment l'intégration de ressources externes (moteurs de recherche, bases de connaissances) dans l'architecture des modèles. Lewis et al. (2020) [6] proposent par exemple des approches de génération ancrée dans la récupération d'information, bien que ces méthodes introduisent des contraintes supplémentaires, notamment en termes de latence et de cohérence globale du discours.

4 Origine et causes des hallucinations

Les hallucinations générées par les LLMs sont le résultat de multiples facteurs interdépendants, allant des données d'entraînement aux limitations architecturales, en passant par les stratégies de génération.

4.1 Qualité des données d'entraînement

Les LLMs apprennent à partir de vastes corpus de textes, souvent collectés automatiquement à partir de sources variées. Lorsque ces données sont incomplètes, biaisées ou contiennent des informations erronées, le modèle peut intégrer et reproduire ces inexactitudes. Par exemple, un modèle entraîné sur des données contenant des informations obsolètes ou des erreurs factuelles est susceptible de générer des contenus incorrects. Ce phénomène est accentué par le principe du "garbage in, garbage out", où la qualité des sorties est directement liée à celle des entrées [5].

4.2 Des modèles limités par leur conception

Les architectures actuelles des LLMs, notamment les transformeurs, ont des capacités limitées à maintenir la cohérence sur de longues séquences. De plus, des biais d'entraînement tels que le "biais d'exposition" (exposure bias) peuvent survenir lorsque le modèle est entraîné à prédire le mot suivant dans une séquence donnée, mais est ensuite utilisé pour générer des séquences entières de manière autonome. Cette divergence entre les phases d'entraînement et d'inférence peut conduire à des incohérences et des hallucinations dans les sorties générées .

4.3 Une génération guidée par la probabilité, pas par la vérité

Les méthodes de génération, telles que l'échantillonnage top-k ou la recherche de faisceaux (beam search), influencent la nature des sorties des LLMs. Ces techniques, bien qu'efficaces pour produire des textes fluides et variés, peuvent également favoriser la génération de contenus plausibles mais incorrects. La nature probabiliste de ces méthodes signifie que le modèle peut produire des informations erronées avec une grande confiance apparente, sans mécanisme intrinsèque de vérification factuelle [16].

4.4 Absence de mécanismes de vérification factuelle

Contrairement aux systèmes dotés de bases de connaissances explicites, les LLMs ne disposent pas de mécanismes intégrés pour vérifier la véracité des informations qu'ils génèrent. Ils se basent sur des corrélations statistiques apprises lors de l'entraînement, sans accès direct à des sources externes pour valider leurs affirmations. Cela les rend susceptibles de produire des informations incorrectes, surtout lorsqu'ils sont confrontés à des requêtes nécessitant une précision factuelle élevée .

4.5 Limites fondamentales des modèles de langage

Des recherches récentes [15] [1] suggèrent que les hallucinations pourraient être une caractéristique inhérente aux LLMs, en raison de leurs fondements mathématiques et logiques. Selon certaines études, il serait impossible d'éliminer complètement les hallucinations, car elles découlent de la manière dont ces modèles apprennent et généralisent à partir des données .

5 Limites structurelles des LLMs et impact sur les hallucinations

Malgré les progrès récents des grands modèles de langage, plusieurs limitations structurelles expliquent pourquoi les hallucinations persistent dans leurs réponses. Ces contraintes affectent la

manière dont les modèles traitent l'information, raisonnent, mémorisent et accèdent aux connaissances, limitant ainsi leur capacité à générer des sorties cohérentes et factuelles.

5.1 Fenêtre de contexte élargie, mais toujours limitée

Les modèles de langage de grande taille (LLMs) ont connu en 2025 une augmentation significative de leur capacité à gérer de longues séquences d'entrée grâce à l'extension de leur fenêtre de contexte. Certains modèles expérimentaux comme LTM-2-Mini développé par Magic.dev peuvent désormais traiter jusqu'à 100 millions de tokens, soit l'équivalent de plus de 10 millions de lignes de code ou environ 750 romans [2]. D'autres modèles comme Claude 2 ou GPT-4-turbo prennent en charge entre 128k et 1M de tokens, selon leur configuration [13].

Malgré ces avancées, cette extension ne résout pas entièrement les problèmes liés à la mémoire effective du modèle. Plusieurs études ont montré que les LLMs n'utilisent pas uniformément toutes les informations présentes dans la fenêtre de contexte : leur attention diminue avec la distance, ce qui peut entraîner une perte d'information sur les tokens plus anciens [8]. Par ailleurs, l'augmentation de la taille du contexte entraîne une croissance quadratique ou supérieure du coût de calcul, selon l'architecture du modèle (Transformer classique vs architectures optimisées), ce qui limite l'usage pratique de ces grandes fenêtres dans des applications courantes [10].

Enfin, une fenêtre de contexte plus large n'implique pas nécessairement une meilleure compréhension, car les LLMs ne possèdent pas de mémoire sémantique stable ni de mécanisme natif de raisonnement logique sur le long terme. Une grande quantité de contexte peut même nuire à la cohérence si le modèle est incapable de hiérarchiser et structurer les informations pertinentes à l'échelle du document.

5.2 Mémoire conversationnelle persistante : une avancée récente

Traditionnellement, les LLMs traitaient chaque requête de manière indépendante, sans mémoire des interactions précédentes. Récemment, des avancées ont permis l'introduction de mémoires conversationnelles persistantes. Par exemple, ChatGPT peut désormais référencer l'historique complet des conversations pour fournir des réponses plus personnalisées [14]. Néanmoins, cette mémoire est encore limitée et nécessite une gestion attentive pour éviter les dérives ou les biais accumulés.

5.3 Accès en temps réel à l'information : progrès et défis

Certains LLMs ont été intégrés à des systèmes permettant un accès en temps réel à des sources d'information externes, comme le web. Cela améliore la pertinence des réponses, notamment pour des événements récents. Cependant, cette capacité introduit également des risques, tels que la propagation de fausses informations ou la dépendance à des sources non vérifiées. Il est donc crucial de combiner ces systèmes avec des mécanismes de vérification et de validation des données.

5.4 Raisonnement causal et logique : des capacités limitées

Malgré leurs performances en génération de texte, les LLMs peinent à effectuer des tâches nécessitant un raisonnement causal ou logique complexe. Par exemple, une étude a révélé que des modèles multimodaux comme GPT-4o et Llama 3.2-Vision échouaient fréquemment à lire correctement des horloges analogiques ou à interpréter des calendriers, avec des taux de réussite respectifs de 38,7% et 26,3% [12]. Ces limitations soulignent la nécessité d'une supervision humaine pour les tâches critiques.

5.5 Biais et sécurité : des préoccupations constantes

Les LLMs peuvent reproduire ou amplifier les biais présents dans leurs données d'entraînement. De plus, des chercheurs ont démontré que la plupart des chatbots IA peuvent être facilement manipulés pour contourner leurs garde-fous de sécurité, fournissant ainsi des informations dangereuses ou illégales [11]. Ces vulnérabilités nécessitent une attention particulière, notamment en matière de filtrage des données, de conception éthique et de régulation.

6 Évaluation et benchmarks

L'évaluation des hallucinations générées par les LLMs reste un défi ouvert dans la recherche actuelle. Les benchmarks traditionnels (comme TruthfulQA, FactScore ou FaithDial) permettent de tester la factualité et la cohérence sur des tâches spécifiques, mais ils souffrent souvent de limites de couverture ou de subjectivité dans l'évaluation humaine des réponses. Par ailleurs, peu de jeux de données sont adaptés à des modèles dotés de longues fenêtres de contexte, ce qui rend difficile l'évaluation à grande échelle sur des documents longs ou des conversations étendues [4].

Les chercheurs soulignent donc la nécessité de développer des outils d'évaluation plus robustes et adaptés aux caractéristiques spécifiques des LLMs modernes, notamment leur capacité à combiner contexte local, mémoire persistante, et raisonnement symbolique. Des approches automatiques fondées sur la vérification par modèles tiers (retrieval-based evaluation) ou la supervision humaine assistée par LLM sont également explorées pour améliorer la détection d'hallucinations.

7 Méthodes de réduction des hallucinations

Face à la persistance des hallucinations, même dans les modèles à fenêtre de contexte étendue ou à mémoire conversationnelle, plusieurs stratégies complémentaires sont explorées.

- **Amélioration des données d'entraînement :** Nettoyer les corpus, diversifier les sources et appliquer des filtres de qualité permet de réduire les biais et incohérences à l'origine des hallucinations. L'usage de données vérifiées et alignées sur des connaissances factuelles est une première ligne de défense [3]. Techniques d'entraînement avancées : Des approches comme le RLHF (Reinforcement Learning from Human Feedback) ou les méthodes contrastives (e.g. DPO – Direct Preference Optimization) permettent d'aligner les sorties du modèle avec les attentes humaines, en privilégiant les réponses précises et utiles [9].
- **Intégration de connaissances externes :** Des modèles hybrides (RAG – Retrieval-Augmented Generation) ou connectés à des bases de connaissances en temps réel (comme WebGPT ou Bing Copilot) peuvent vérifier et compléter leurs sorties avec des informations factuelles, réduisant les hallucinations extrinsèques [6].
- **Fenêtres de contexte élargies et mémoire persistante :** L'élargissement du contexte, comme observé avec Claude 2 ou GPT-4-turbo (jusqu'à 1 million de tokens), permet d'intégrer davantage d'informations en entrée. Toutefois, des travaux montrent que cette capacité est limitée par l'attention décroissante sur les tokens éloignés, d'où l'intérêt croissant pour des architectures capables de mieux structurer ou hiérarchiser l'information [10].
- **Surveillance et redétection en temps réel :** Certains systèmes récents intègrent une vérification automatique des réponses générées en temps réel, avec rejet ou reformulation en cas d'erreur détectée. Ce type de pipeline est notamment exploré dans les assistants IA déployés à grande échelle (e.g. chez Anthropic ou OpenAI).

8 Perspectives et solutions

Pour atténuer les effets des hallucinations, plusieurs pistes sont actuellement explorées par la communauté scientifique.

Une première approche consiste à connecter les LLMs à des sources d'information externes fiables, comme des bases de données ou des API (par exemple, Bing Search ou WolframAlpha). Cela permettrait de vérifier les faits en temps réel et de réduire les erreurs factuelles.

Une autre solution prometteuse repose sur l'apprentissage par renforcement avec retour humain (RLHF), qui permet d'aligner les sorties du modèle avec les attentes humaines. Cette méthode, déjà utilisée pour affiner ChatGPT, améliore la pertinence des réponses, bien qu'elle ne garantisse pas l'absence d'erreurs.

Enfin, des techniques de détection automatique d'hallucinations sont en cours de développement. Celles-ci visent à identifier, en sortie de modèle, les passages douteux ou incorrects en les comparant à des sources ou à l'aide d'un second modèle de vérification.

Il est probable que la combinaison de plusieurs techniques soit nécessaire pour garantir une fiabilité suffisante, en particulier dans les applications critiques.

9 Conclusion

Les hallucinations générées par les grands modèles de langage illustrent les limites fondamentales de ces systèmes statistiques. Malgré leurs performances impressionnantes, les LLMs restent sensibles à la qualité de leurs données d'entraînement et à la structure probabiliste de leur génération de texte, ce qui les rend vulnérables à la production de contenus erronés.

La compréhension des causes profondes de ces hallucinations, ainsi que le développement de métriques fiables pour les détecter, sont des étapes cruciales vers une utilisation plus responsable de ces technologies. Les solutions proposées (ancrage dans des sources externes, apprentissage avec retour humain, ou détection automatique) offrent des perspectives encourageantes, mais soulignent aussi la complexité du défi à relever.

À mesure que les LLMs s'intègrent dans des domaines sensibles, il devient impératif de renforcer leur fiabilité et leur transparence, pour que leur puissance ne soit pas synonyme de danger.

Références

[1] Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv :2409.05746*, 2024.

[2] Codingscape. Llms with the largest context windows, 2024.

[3] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017) :625–630, 2024.

[4] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models : Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2) :1–55, 2025.

[5] Sajjad Khazipura. Why do large language models hallucinate ?, May 2025.

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33 :9459–9474, 2020.

[7] Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, Li Zhang, Zhongqi Li, and Yuchi Ma. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv :2404.00971*, 2024.

[8] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle : How language models use long contexts. *arXiv preprint arXiv :2307.03172*, 2023.

[9] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35 :27730–27744, 2022.

[10] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long : Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv :2108.12409*, 2021.

[11] Ian Sample. Most ai chatbots easily tricked into giving dangerous responses, study finds, May 2025. Consulté le 29 mai 2025.

[12] Drew Turney. Ai models can't tell time or read a calendar, study reveals, May 2025. Consulté le 29 mai 2025.

[13] Akihiko Wada, Toshiaki Akashi, George Shih, Akifumi Hagiwara, Mitsuo Nishizawa, Yayoi Hayakawa, Junko Kikuta, Keigo Shimoji, Katsuhiro Sano, Koji Kamagata, et al. Optimizing gpt-4 turbo diagnostic accuracy in neuroradiology through prompt engineering and confidence thresholds. *Diagnostics*, 14(14) :1541, 2024.

[14] Simon Willison. I really don't like chatgpt's new memory dossier, May 2025.

[15] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable : An innate limitation of large language models. *arXiv preprint arXiv :2401.11817*, 2024.

[16] Zep. Reducing llm hallucinations : A developer's guide, 2024.

[17] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean : a survey on hallucination in large language models. *arXiv preprint arXiv :2309.01219*, 2023.