



HAL
open science

Une revue sur les hallucinations des LLM

Eleni Metheniti, Swarnadeep Bhar, Nicholas Asher

► **To cite this version:**

Eleni Metheniti, Swarnadeep Bhar, Nicholas Asher. Une revue sur les hallucinations des LLM. 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux (CORIA-TALN 2025), Jun 2025, Marseille, France. <hal-05279740>

HAL Id: hal-05279740

<https://hal.science/hal-05279740v1>

Submitted on 24 Sep 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Une revue sur les hallucinations des LLM

Eleni Metheniti ¹ Swarnadeep Bhar ² Nicholas Asher ^{2,3}

(1) Artificial and Natural Intelligence Toulouse Institute (ANITI), 31400 Toulouse, France

(2) Institut de Recherche en Informatique de Toulouse (IRIT), 31400 Toulouse, France

(3) CNRS, 31400 Toulouse, France

`eleni.metheniti@univ-tlse3.fr`

RÉSUMÉ

Nous présentons une taxonomie des hallucinations dans les LLM, classées en trois catégories : hallucinations infidèles, contradictions factuelles et fabrications factuelles. Ces hallucinations peuvent se produire à cause des données de pré-entraînement et d'alignement, conduisant à des informations erronées, des préjugés et des erreurs de connaissance. Les méthodes d'entraînement peuvent introduire des problèmes tels que l'ajustement excessif, les effets boule de neige ou la sycophantie. Les stratégies de décodage peuvent également rendre les modèles trop confiants et enclins à attribuer des probabilités aux résultats incorrects. Une bibliographie sur la détection et atténuation des hallucinations est présentée : des méthodes de TALN, telles que la vérification des faits et la classification, de même que des méthodes basées sur les LLM. Les solutions d'atténuation des hallucinations comprennent l'amélioration de la qualité des données de pré-entraînement, l'injection de nouvelles connaissances (par ex. avec RAG), l'optimisation, SFT et RLHF, ainsi que des méthodes de décodage.

ABSTRACT

A survey on LLM hallucinations

We present a taxonomy of hallucinations in LLMs, categorized into three types : unfaithful hallucinations, factual contradictions, and factual fabrications. These hallucinations can be caused by pretraining and alignment data, leading to misinformation, biases, and lapses in knowledge. Training methods can introduce problems like overfitting, snowball effects, or sycophancy. Decoding strategies can also render models overconfident and prone to assigning probabilities to incorrect outputs. We present current research on detecting and mitigating hallucinations ; traditional NLP methods, such as fact-checking and classification, as well as LLM-based methods. Mitigation solutions include improving pretraining data quality, injecting new knowledge such as with RAG, optimizing methods, SFT, and RLHF, and factuality and post-editing decoding methods to improve LLM output.

MOTS-CLÉS : hallucinations des LLM, détection des hallucinations, atténuation des hallucinations.

KEYWORDS: LLM hallucinations, hallucination detection, hallucination mitigation.

ARTICLE : **Soumis à TALN 2025.**

1 Introduction

L'introduction de *transformers* (Vaswani *et al.*, 2017) a ouvert la voie à la création de grands modèles de langage (*large language models*; LLMs), c'est-à-dire des modèles qui sont entraînés sur de grande quantité de données textuelles, utilisant généralement le pré-entraînement et l'auto-supervision, contenant au moins 1M de paramètres et applicables dans un large éventail de contextes. Les LLMs ont révolutionné le domaine du traitement du langage naturel (NLP), établissant un nouvel état de l'art sur des tâches telles que la génération de langage naturel (Dugan *et al.*, 2023; Zhang *et al.*, 2024a) et le raisonnement (Qiao *et al.*, 2022; Yu *et al.*, 2024). Leurs capacités avancées peuvent être attribuées à leurs architectures sous-jacentes de *transformers* et d'autorégression, ainsi qu'à des quantités massives de données (corpus de textes ou données multimodales) entraînées avec une infrastructure informatique puissante.

Malgré leur adaptation répandue aux tâches du TALN, les LLMs se heurtent à des problèmes de fiabilité et de raisonnement avancé sur les connaissances acquises. Ils sont enclins à générer des textes qui peuvent sembler plausibles mais qui peuvent s'écarter du contexte précédent, des instructions de l'utilisateur ou des connaissances factuelles. Ces informations inexacts ou fabriquées sont communément appelées «hallucinations». Les hallucinations peuvent s'expliquer par le volume massif de données d'entraînement, l'architecture sous-jacente des modèles ou la nature polyvalente des LLMs qui favorise les performances inter-domaines au détriment d'une compréhension et d'une connaissance approfondies (Zhang *et al.*, 2023d; Rawte *et al.*, 2023; Huang *et al.*, 2025). Par ailleurs, la recherche sur l'atténuation des hallucinations est en plein essor (Maynez *et al.*, 2020; Liu *et al.*, 2023; Tonmoy *et al.*, 2024). Dans cet article, nous présentons une taxonomie des différents types d'hallucinations qui se produisent dans les sorties des LLMs (Section 2). Nous explorons les causes possibles de ces hallucinations pour chaque étape de la création d'un LLM dans la Section 3. Enfin, nous examinons les méthodes pour détecter (Section 4) et atténuer (Section 5) les hallucinations dans les sorties des LLMs

2 Taxonomie

Le terme «hallucination» a été inspiré par l'homonyme psychiatrique préexistant; selon Boksa (2009), les hallucinations chez les humains sont des perceptions qui se produisent sans déclencheur externe et sont renforcées par une forte sensation de réalité. Dans le domaine du TALN, et plus particulièrement de la génération de langage naturel, les chercheurs ont commencé à appeler familièrement «hallucinations» les sorties de modèles contenant des faits inventés, des informations infidèles et des déclarations absurdes; ce terme a été largement adopté ces dernières années (Zhou *et al.*, 2021; Li *et al.*, 2024b). L'exactitude grammaticale et la fluidité du texte généré peuvent rendre les hallucinations difficiles à détecter, en particulier en ce qui concerne les faits et les informations historiques. Ces résultats invérifiables ou incohérents posent un problème et constituent potentiellement une menace, à la fois pour l'avenir de l'intelligence artificielle et pour les destinataires humains de ces résultats.

Selon les différents types d'erreurs produites, une taxonomie des hallucinations, basée sur les travaux de Huang *et al.* (2025) et Li *et al.* (2024b) peut être définie comme suit. Des exemples de chaque type d'hallucination, produits par divers modèles, sont présentés en Annexe.

Hallucinations de fidélité :

- Incohérences d'instruction
- Incohérences de contexte
- Incohérences logiques

Erreurs de faits :

- Erreurs d'entité
- Erreurs de relation
- Erreurs d'incomplétude
- Erreurs d'obsolescence

Fabrication de faits :

- Hallucinations invérifiables
- Hallucinations de surproclamation

2.1 Hallucinations de fidélité

Les hallucinations de fidélité se produisent lorsque le texte généré par le modèle ne correspond pas aux instructions données par l'utilisateur, au contexte de l'entrée ou au texte généré précédemment. Cela signifie que la réponse du modèle n'a finalement aucun rapport avec la requête de l'utilisateur ou qu'elle peut être illogique. Ces hallucinations peuvent être divisées en trois types d'incohérences :

Les **incohérences d'instruction** se produisent lorsque la sortie du modèle s'écarte des instructions de l'utilisateur. Il peut s'agir d'une caractéristique de conception intentionnelle d'un chatbot commercial alimenté par LLM, par exemple, afin d'éviter toute utilisation malveillante. Cependant, il n'est pas rare que le modèle manque des indices dans l'entrée de l'utilisateur et dévie des instructions données. Un exemple peut être vu dans le Tableau 1.

Les **incohérences de contexte** sont les cas où la sortie générée par le modèle ne correspond pas au contexte fourni par l'utilisateur. Dans l'exemple du Tableau 2, le modèle ignore les instructions de l'utilisateur et produit une recette avec des ingrédients qu'il n'a pas.

Les **incohérences logiques** sont des erreurs qui se produisent lorsque la sortie du modèle ne suit pas les étapes logiques de son raisonnement. Par exemple, les erreurs de calcul peuvent être classées dans la catégorie des incohérences logiques, car les modèles ne parviennent pas toujours à analyser et à suivre toutes les étapes nécessaires à la résolution des équations. Dans le Tableau 3, le modèle a généré une hallucination en donnant une réponse incorrecte aux calculs, alors qu'il avait déjà trouvé la bonne réponse. Concernant ce type d'hallucinations, [Asher & Bhar \(2024\)](#) les caractérisent comme *hallucinations fortes* lorsque le modèle fournit comme preuve de son raisonnement des affirmations qui sont logiquement insatisfaisantes ou contradictoires.

2.2 Erreurs de faits

Les **erreurs de faits** se produisent quand la sortie du modèle contient des faits du monde réel qui se contredisent. Il s'agit de la forme d'hallucination la plus courante, qui peut être due à divers facteurs, notamment la capacité du LLM à enregistrer, à stocker et à communiquer des connaissances exactes. Deux sous-catégories peuvent être identifiées en fonction du type d'incohérence qui se produit : l'hallucination par erreur d'entité et l'hallucination par erreur de relation.

Les **erreurs d'entité** décrivent les cas où le texte produit par les LLMs contient des entités incorrectes, telles qu'une personne, une date, un lieu ou un objet, qui sont en contradiction avec les connaissances du monde. Par exemple, dans le Tableau 4, le modèle associe des noms de personnages historiques réels à des dates de naissance ou de décès, des professions ou faits incorrects.

Les **erreurs de relation** surviennent lorsque la sortie du modèle inclut des relations incorrectes entre les faits, telles que des relations quantitatives ou temporelles. Par exemple, dans le Tableau 5, le modèle est interrogé sur les médailles olympiques d'un athlète et lui attribue une médaille d'argent

qu'il n'a pas remportée lors de cet événement spécifique.

Les **erreurs d'incomplétude** se produisent lorsque la sortie du LLM contient des informations incomplètes lors de la génération de réponses longues ou énumérées. Lorsque les LLMs sont interrogés sur des informations agrégées, ils peuvent générer des hallucinations en raison d'une incertitude quant à la véracité des faits. Dans le Tableau 6, par exemple, le modèle ne parvient pas à récupérer les informations complètes sur les billets pour tous les sites touristiques énumérés dans la réponse qu'il a générée.

Les **erreurs d'obsolescence** se produisent lorsque les résultats du modèle mentionnent un fait qui n'est plus correct parce que l'ensemble des données d'entraînement contient des informations qui ne sont plus d'actualité. Par exemple, dans le Tableau 7, la réponse du modèle à la question de savoir qui est le Premier ministre de la France est obsolète depuis plus d'un an.

2.3 Fabrication de faits

La fabrication de faits se produit lorsque le modèle invente un fait qui ne s'est jamais produit dans le monde réel. Si ces inventions sont plausibles, et si l'utilisateur n'est pas familiarisé avec le sujet ou les faits en question, il peut être amené à les croire et à les reproduire.

Les **hallucinations invérifiables** se trouvent dans les résultats qui contiennent de nombreuses informations non factuelles, inventées et parfois absurdes. Les faits générés ne peuvent être vérifiés à partir d'aucune source connue, ce qui indique qu'il s'agit d'une invention. Par exemple, dans le Tableau 8, le modèle hallucine longuement sur l'emplacement et l'histoire d'un site célèbre, invente des personnages et des faits inexistantes et confond des époques historiques.

On parle d'**hallucinations de surproclamation** lorsque les faits générés parlent d'un événement ou d'une occurrence qui est plausible, mais il n'y a aucune preuve de son existence. Par exemple, dans le Tableau 9, le modèle génère des faits exagérés lorsqu'on lui demande de créer un résumé de sa réponse précédente contenant des faits correctes.

3 Causes des hallucinations par étape du LLM

3.1 Données de pré-entraînement

Les LLMs nécessitent de grandes quantités de données de pré-entraînement ; par exemple, GPT-3 (Brown *et al.*, 2020) a été entraîné sur 45 To de texte brut compressé (570 Go après filtrage) provenant de Wikipédia, d'Internet et de livres. Ces données constituent la base de la connaissance des LLM, des sujets et des faits qu'ils stockent ; combinés à de puissantes architectures neuronales telles que les transformers, les LLMs peuvent acquérir une connaissance approfondie du monde et créer des connexions saillantes entre les informations. Cependant, ça laisse les modèles vulnérables aux erreurs (Lin *et al.*, 2022) et biais (Ladhak *et al.*, 2023), et l'énorme volume de données textuelles rend leur vérification manuelle impossible. Par exemple, les modèles peuvent reproduire des informations erronées présentes dans les données (Lin *et al.*, 2022; Naim *et al.*, 2024) ou répéter des stéréotypes sociétaux nuisibles en raison de leur fréquence élevée (Ladhak *et al.*, 2023). Cependant, Jiang *et al.* (2024) soutiennent qu'il n'y a pas de corrélation significative entre la popularité d'un sujet dans les données existantes et les erreurs d'entités.

Un autre défi est l'obsolescence ; sans l'ajout régulier de nouvelles données, les modèles n'ont accès qu'aux faits existant au moment où les données ont été créées (Jiang *et al.*, 2024). Ainsi, les LLMs inventent souvent des faits ou donnent des réponses qui ont pu être exactes dans le passé mais qui sont désormais dépassées lorsqu'ils sont confrontés à des questions postérieures à leurs données d'entraînement (Onoe *et al.*, 2022).

De plus, les LLMs présentent différents degrés de compétence dans divers domaines de connaissance en raison de la distribution naturellement inégale des informations dans les données de pré-entraînement. Selon Raffel *et al.* (2020); Kandpal *et al.* (2023), il existe une corrélation entre la quantité de contenus pertinents dans les corpus de pré-entraînement et la précision des modèles sur les requêtes du domaine général. En particulier dans les tâches complexes de TALN, telles que le raisonnement et l'ILN (Inférence en Langage Naturel), McKenna *et al.* (2023); Naim & Asher (2025) observent que la fréquence des entités dans les données de pré-entraînement est un meilleur indicateur pour prédire l'implication que les déductions à partir du contexte.

3.2 Pré-entraînement

L'étape de pré-entraînement des LLMs génératifs utilise un objectif de *modélisation causale du langage*, dans lequel les modèles sont entraînés sur le contexte passé pour prédire les nouveaux tokens de manière unidirectionnelle. Bien que cette méthode rende l'entraînement plus efficace, elle limite naturellement la capacité à identifier des relations contextuelles complexes, ce qui peut augmenter la probabilité d'hallucinations. Des éléments clés des modèles, tels que l'objectif de l'entraînement autorégressif (Asher & Bhar, 2024) et le mécanisme d'attention (Chiang & Cholak, 2022; Naim & Asher, 2024), pourraient également conduire à des hallucinations infidèles. Les LLMs acquièrent des connaissances factuelles en accumulant des micro-acquisitions et en les oubliant par la suite.

Selon Chang *et al.* (2024), il existe en théorie un *seuil d'apprenabilité*, c'est-à-dire une limite de rencontres après laquelle le modèle ne parviendra pas à acquérir de nouvelles connaissances. Les auteurs affirment que, pour la plupart des faits bien connus, leur occurrence dans les données d'apprentissage est inférieure à ce seuil, mais pourrait potentiellement conduire à des hallucinations factuelles. Jiang *et al.* (2024) examinent l'état interne de LLaMA-2 (Touvron *et al.*, 2023), ses représentations abstraites ou sémantiques apprises de tokens, et observent que les hallucinations factuelles peuvent se produire sur les couches finales de la perception multicouche, contribuant à un décodage de sortie erroné. Enfin, l'utilisation d'algorithmes de décodage variables, en tant que méthode pour produire un caractère aléatoire dans le texte généré, peut produire une plus grande diversité mais aussi des hallucinations (Brown *et al.*, 2020; Zhang *et al.*, 2021).

D'un point de vue mathématique, Xu *et al.* (2024b) théorisent que les hallucinations dans les LLMs sont inhérentes, indépendamment de l'architecture du modèle, des algorithmes d'apprentissage, des données d'entraînement et d'autres détails de mise en œuvre. Ils prouvent mathématiquement que les LLMs à temps polynomial hallucineront sur des problèmes du monde réel, par exemple l'arithmétique, et ce, sur un nombre infini d'entrées. En outre, Asher & Bhar (2024) proposent également une preuve mathématique que la façon d'attribuer des probabilités aux nouvelles chaînes dans un modèle génératif conduira *de facto* à ce que des distributions non objectives deviennent plausibles.

3.3 Ajustement supervisé

L'ajustement supervisé (*Supervised Fine-tuning* ; SFT) est utilisé pour former davantage les LLMs afin d'améliorer leurs capacités (Chung *et al.*, 2024). Bien que ce processus soit considéré comme une méthode relativement stable, il court également le risque de *surapprentissage* avec de nouvelles connaissances factuelles, produisant ainsi des hallucinations factuelles. Pendant le SFT, il est demandé au modèle de générer des faits qui ne sont pas fondés sur ses connaissances préexistantes ; le modèle doit donc apprendre ces faits uniquement via SFT (Gekhman *et al.*, 2024). De plus, lorsque l'on demandera au modèle de produire un résultat, il s'efforcera de fournir une réponse, quelle qu'en soit la véracité. (Zhang *et al.*, 2023a).

3.4 Apprentissage par renforcement à partir de rétroaction humaine

L'apprentissage par renforcement à partir de rétroaction humaine (*Reinforcement learning from human feedback* ; RLHF) utilise le *feedback* (retour) humain pour évaluer les résultats des LLMs et entraîne ensuite un modèle de récompense à partir de ce feedback, afin d'aligner le modèle sur les croyances humaines. Il a été observé que les LLMs ont tendance à halluciner afin de s'aligner sur le feedback humain plutôt que de donner des réponses fondées sur leurs croyances apprises (Perez *et al.*, 2022). L'instabilité pendant l'optimisation RL peut rendre le LLM sensible aux inexactitudes du modèle de récompense, ce qui peut entraîner un comportement inattendu. La qualité des données est cruciale pour le RLHF, qui repose sur un retour d'information humain précis. En outre, le RLHF spécifique à un domaine peut désavantager les LLMs du domaine général (Li *et al.*, 2024b).

4 Détecter les hallucinations

4.1 Métriques basées sur les faits

L'*extraction de faits* sert à extraire les déclarations factuelles faites dans les résultats du LLM. Elle peut être réalisée à l'aide d'outils externes, par exemple, des mesures qui décomposent le texte généré en faits et les comparent à des sources de connaissances externes (Min *et al.*, 2023; Chern *et al.*, 2023; Wang *et al.*, 2024b), ou en vérifiant la sortie en interrogeant le modèle avec une combinaison de l'entrée de l'utilisateur et de la sortie (Huo *et al.*, 2023). Les mesures *n-gramme* traditionnelles ne parviennent pas à distinguer les différences subtiles entre les informations produites et celles issues des sources (Maynez *et al.*, 2020). Afin de détecter les associations erronées ou l'infidélité au contexte, les mesures basées sur les entités mesurent le chevauchement des entités (Goodrich *et al.*, 2019) ou le chevauchement des tuples de relations extraits à l'aide de modèles d'extraction de faits formés de bout en bout (Nan *et al.*, 2021).

4.2 Métriques basées sur la classification

Pour évaluer la fidélité des résultats générés par le LLM, il est possible d'utiliser des méthodes de classification de tâches du TALN, par exemple, l'inférence du langage naturel (Rahimi *et al.*, 2024). Cependant, les mesures traditionnelles du TALN n'ont pas été conçues pour détecter le type d'hallucinations qui se produisent avec les LLMs (Mishra *et al.*, 2021); de plus, elles échouent sur les langues à faibles ressources (Kang *et al.*, 2024). Ainsi, ces mesures ne peuvent évaluer que des

aspects spécifiques de la sortie d'un modèle, et ne saisissent donc pas nécessairement tous les types d'hallucinations possibles dans un énoncé (Bang *et al.*, 2023).

4.3 Métriques basées sur la question-réponse

Les métriques basées sur la question-réponse (*Question-Answering*; QA) calculent le taux de réponses correctes dans les sorties du modèle (Lin *et al.*, 2022; Min *et al.*, 2023; Muhlgay *et al.*, 2024; Wang *et al.*, 2024c; Li *et al.*, 2024b). Ces mesures choisissent des réponses cibles à partir des résultats du modèle, produisent des questions avec le module de génération de questions et évaluent la fidélité des réponses en comparant les scores de correspondance entre les réponses source et cible. La vérification de ces informations peut être effectuée à l'origine par des humains (Zhao *et al.*, 2024).

4.4 Estimation de l'incertitude

L'estimation de l'incertitude se réfère aux états internes du modèle et peut donc être utilisée sans besoin d'outils externes. La probabilité logarithmique de la séquence prédite peut être utilisée à plusieurs niveaux de granularité, comme le niveau du mot ou de la phrase (Xiao & Wang, 2021; Varshney *et al.*, 2023; Guerreiro *et al.*, 2023; Fu *et al.*, 2024). Luo *et al.* (2024) demandent au modèle de générer une explication du concept prédit via un décodage contraint, afin de recréer la réponse originale sur la base de l'explication générée. Les méthodes adverses peuvent également être utilisées pour provoquer des hallucinations et ainsi étudier les états du modèles (probabilité de jeton, entropie, etc.) puissent être étudiés (Yao *et al.*, 2023). Zhang *et al.* (2023c) ont mis au point un système de sanction pour empêcher la diffusion de documents hallucinés en se concentrant sur les termes les plus informatifs et les plus significatifs.

4.5 Métriques basées sur les LLM

Il est possible d'utiliser les **LLM eux-mêmes** comme source de vérification des faits. Par exemple, Dhuliawala *et al.* (2024) utilisent un LLM pour générer des questions de vérification à partir de la réponse du modèle, pour répondre aux questions générées et pour comparer la cohérence de la réponse originale et de la nouvelle réponse. De même, Li *et al.* (2024b) demandent au LLM d'extraire de la réponse des déclarations qui ressemblent à des faits, puis utilisent un autre LLM pour vérifier leur véracité. La recherche a également tenté de fournir aux LLM des directives d'évaluation, puis de demander aux modèles d'effectuer une auto-évaluation de leur propre résultat (Luo *et al.*, 2023; Laban *et al.*, 2023). En ce qui concerne la certitude d'un modèle sur la qualité de sa production, Manakul *et al.* (2023) ont identifié les hallucinations en évaluant la cohérence des assertions vraies.

Cependant, ces stratégies sont principalement basées sur des requêtes directes qui recherchent des informations ou des vérifications auprès du modèle (Kadavath *et al.*, 2022; Xiong *et al.*, 2023). Agrawal *et al.* (2024) ont plaidé en faveur des enquêtes indirectes, qui posent généralement des questions ouvertes pour extraire des informations particulières. Cohen *et al.* (2023) utilisent un LLM examinateur pour interroger un autre LLM, afin de trouver des divergences pendant un dialogue à plusieurs tours.

5 Atténuer les hallucinations

5.1 Atténuer les hallucinations liées aux données

5.1.1 Filtrage des données

Les données scientifiques et des domaines spécialisés sont généralement fiables et sans biais forts, elles sont donc des ressources convoitées pour l'entraînement (Gao *et al.*, 2020; Gunasekar *et al.*, 2023). Cela garantit l'exactitude des faits et minimise l'introduction de préjugés sociaux. En outre, Touvron *et al.* (2023) ont démontré que le *suréchantillonnage* des données factuelles au cours du pré-entraînement, c'est-à-dire l'ajout d'échantillons originaux de classes minoritaires jusqu'à que les classes soient équilibrées, améliore l'exactitude des faits et limite les hallucinations. La déduplication peut également améliorer la qualité des données en éliminant les jetons de données répétitives. Pour les jetons exacts, la méthode la plus simple consiste à faire correspondre des sous-chaînes exactes, mais cette méthode peut s'avérer très coûteuse en raison de l'étendue des données de préapprentissage. Les quasi-jetons impliquent souvent une correspondance approximative en texte intégral ou des jetons sémantiques, généralement à l'aide de techniques basées sur le hachage (Gyawali *et al.*, 2020; Abbas *et al.*, 2023).

5.1.2 Révision du modèle

La révision du modèle est une technique utilisée pour améliorer le comportement du modèle en incorporant des connaissances supplémentaires (Wang *et al.*, 2024c). Les méthodes impliquent la localisation d'une partie des paramètres du modèle susceptible de produire des hallucinations, puis l'application d'une mise à jour pour modifier le comportement du modèle (Meng *et al.*, 2022a). Par exemple, Meng *et al.* (2022b) permettent la mise à jour simultanée de plusieurs couches afin d'intégrer des milliers de connaissances de correction. Toutefois, ces méthodes manquent de capacités de généralisation et varient en termes de performances et d'applicabilité à différentes architectures de modèles.

D'autre part, la technique de *meta-learning* entraîne un hyper-réseau externe à prédire la mise à jour du poids du modèle original (De Cao *et al.*, 2021). Cependant, ces méthodes nécessitent souvent des coûts supplémentaires d'entraînement et de mémoire, et les modifications des paramètres peuvent potentiellement nuire à la connaissance inhérente du modèle, en particulier par des changements continus. Tan *et al.* (2023), par exemple, ont abordé ce problème en réduisant l'agrégation des changements de paramètres.

5.1.3 Graphes de connaissances

Les graphes de connaissances (*Knowledge Graphs*; KG) sont des bases de connaissances qui utilisent un modèle de données structuré en graphe ou une topologie pour représenter et exploiter les données. Ils stockent des descriptions interconnectées d'entités ainsi que les relations et la sémantique qui les unissent. L'exploitation de ces bases de connaissances permet d'intégrer des connaissances spécialisées ou actualisées et de favoriser de raisonnement solides (Baek *et al.*, 2023; Wen *et al.*, 2024; Varshney *et al.*, 2023). Leur structure codifie les relations entre les entités sous forme de triplets de connaissances : $\langle head, relation, tail \rangle$, où *head* et *tail* sont des entités reliées par une *relation* (Rossi *et al.*, 2021). La distance entre les *embeddings* des entités de tête et de queue quantifie la fiabilité d'un fait (Alshahrani *et al.*, 2021). Les KG peuvent être incorporés dans un LLM en tant

que *knowledge graph embeddings* pendant le processus d'entraînement ou d'inférence ; ils peuvent être intégrés en tant qu'embeddings supplémentaires ou en tant que connaissances structurées dans un graphe. Les KG peuvent être utilisés pour l'ajustement des instructions d'un LLM, c'est-à-dire l'ajustement pour suivre les faits et la structure d'un KG afin de mieux exploiter les connaissances. Ils peuvent aussi être intégrés dans la phase d'inférence, avec la RAG ou avec du *prompting* avec des graphes de connaissances (Wen *et al.*, 2024).

5.1.4 Génération augmentée de récupération

La génération augmentée de récupération (*Retrieval-Augmented Generation* ; RAG) est une méthode qui utilise un pipeline de récupération puis de lecture pour extraire les connaissances pertinentes de sources externes et les ajouter à la sortie du LLM, afin d'étendre les connaissances du modèle et de faciliter son application dans divers domaines (Guu *et al.*, 2020). Cependant, la RAG peut être affectée par des extractions non pertinentes et des interactions peu profondes entre les composants de l'extracteur et du générateur.

La **récupération ponctuelle** injecte directement des connaissances externes dans le *prompt* (Ram *et al.*, 2023). Comme indiqué précédemment, l'utilisation de graphes de connaissances s'est également répandue (Pan *et al.*, 2024). Varshney *et al.* (2023) ont aussi créé un cadre paramétrique afin d'injecter des connaissances spécifiques à la tâche dans le LLM.

La **récupération itérative** est une méthode qui permet une collecte continue des connaissances tout au long du processus de génération, ce qui effectue de relever des défis complexes tels que le raisonnement en plusieurs étapes et la réponse à des questions de longue distance (Wei *et al.*, 2022). Press *et al.* (2023) introduisent des questions *multi-hop* des réponses nécessitant la composition de faits assez peu observés ensemble durant l'entraînement. Feng *et al.* (2024) ont utilisé un cadre collaboratif de RAG, où la réponse d'un modèle sert de contexte pertinent pour récupérer des connaissances plus pertinentes. Jiang *et al.* (2023) ont proposé un cadre de RAG, traitant les prédictions à venir comme des requêtes pour récupérer des documents pertinents. Zhang *et al.* (2024b) affinent de manière itérative les questions des utilisateurs à l'aide de conseils basés sur des modèles et demandent des clarifications aux utilisateurs.

La **récupération post hoc** implique la recherche d'informations pertinentes dans les résultats générés et la révision des résultats initiaux en cas d'erreurs (Gao *et al.*, 2023) ou l'incorporation directe de connaissances externes (Zhao *et al.*, 2023). Yu *et al.* (2023) améliorent encore la méthode en échantillonnant diverses réponses potentielles et en employant une technique d'assemblage pour atténuer le risque d'un résultat erroné.

5.2 Atténuer les hallucinations liées à l'entraînement

5.2.1 Hallucinations liées au pré-entraînement

La recherche sur la réduction des hallucinations liées au pré-entraînement s'est concentrée sur les limites des architectures de modèles, en particulier leur mono-directionnalité et le mécanisme d'auto-attention. Des modèles *encoder-decoder* ont aussi été explorés pour un meilleur usage des fenêtres contextuelles, suggérant des perspectives prometteuses pour les futures architectures (Liu *et al.*, 2024). Des techniques d'orientation de l'attention ont également été proposées afin de disperser et réduire les hallucinations du raisonnement (Zhang *et al.*, 2023b).

Pendant la phase de pré-entraînement, le choix de l'objectif joue un rôle crucial dans l'apparition des informations fragmentées et des associations d'entités incorrectes. [Lee et al. \(2022\)](#) ont introduit une méthode de formation par échantillonnage de noyaux factuels, adaptant dynamiquement leur probabilité, afin d'améliorer la factualité tout en maintenant la qualité et la diversité. [Shi et al. \(2023\)](#) ont proposé un pré-entraînement contextuel, qui entraîne les LLMs sur des séquences de documents connexes afin de maximiser la similarité dans les fenêtres contextuelles.

5.2.2 Hallucinations liées à l'alignement

Les hallucinations pendant l'alignement peuvent être causées par un mauvais alignement des capacités et des croyances, qui est difficile à définir dans les LLMs. Ces désalignements peuvent conduire à la *sycophantie*, où les LLMs cherchent l'approbation humaine de manière indésirable ([Sharma et al., 2024](#)). L'une des raisons possibles de ce comportement peut provenir de l'étape du jugement humain lorsque les annotateurs humains favorisent les réponses flatteuses par rapport aux réponses véridiques. Pour y remédier, les chercheurs ont exploré l'utilisation des LLMs pour aider les annotateurs humains à identifier les défauts négligés ([Saunders et al., 2022](#)). [Wei et al. \(2023\)](#) ont proposé d'ajouter des données synthétiques pour affiner les modèles, afin de réduire les tendances à la sycophantie. Une autre approche est l'orientation de l'activation, qui implique l'utilisation de paires de sorties sycophantiques/non sycophantiques afin de générer un vecteur d'orientation ; soustraire ce vecteur au cours de l'inférence peut réduire les sorties sycophantiques ([Panickssery et al., 2023](#)).

5.3 Atténuer les hallucinations liées au décodage

5.3.1 Décodage de la factualité

Le décodage de la factualité est une stratégie utilisée dans l'apprentissage automatique pour trouver un équilibre entre la production de contenu factuel et la préservation de la diversité des résultats. Comme indiqué, [Lee et al. \(2022\)](#) ajustent dynamiquement la probabilité du noyau pour chaque phrase dans la sortie générée en fonction des facteurs de décroissance et des limites inférieures. *L'intervention au moment de l'inférence* est une autre méthode qui identifie une direction dans l'espace d'activation associé aux déclarations factuellement correctes et ajuste les activations dans la direction corrélée à la vérité pendant l'inférence ([Li et al., 2024c](#)). [Chuang et al. \(2024\)](#) explorent la factualité des LLMs en utilisant l'encodage hiérarchique de la connaissance factuelle dans les LLM. Ils emploient une stratégie qui sélectionne et oppose dynamiquement les *logits* de différentes couches en privilégiant les connaissances des couches supérieures et en défavorisant celles des couches inférieures.

5.3.2 Décodage post-édition

Le décodage post-édition utilise les capacités d'autocorrection des LLMs pour affiner le contenu sans s'appuyer sur des connaissances externes. La *chaîne de vérification* et les processus itératifs d'autoréflexion sont utilisés pour s'assurer que les LLMs peuvent fournir des faits plus précis en révisant leurs réponses ([Dhuliawala et al., 2024](#)). [Ji et al. \(2023\)](#) introduisent un processus d'autoréflexion itératif qui intègre l'acquisition de connaissances et la génération de réponses et aligne progressivement les résultats du LLM sur les connaissances de base fournies.

5.3.3 Consistance du contexte

En ce qui concerne les LLMs, le problème des hallucinations de fidélité découle souvent d'une attention insuffisante portée au contexte donné. Cela a conduit à l'élaboration de stratégies visant à améliorer la cohérence contextuelle. Le décodage tenant compte du contexte (CAD) est une méthode qui modifie la distribution de sortie du modèle, lui permettant de se concentrer davantage sur les informations contextuelles (Shi *et al.*, 2023).

Toutefois, accorder trop d'importance aux informations contextuelles peut réduire la diversité. Chang *et al.* (2023) ont présenté un algorithme de décodage dynamique pour maintenir la fidélité tout en préservant la diversité, en utilisant un signal de guidage pour indiquer la pertinence du contexte de la source. Choi *et al.* (2023) ont introduit le **décodage contraint par la connaissance**, qui utilise un discriminateur de détection d'hallucinations au niveau du jeton pour identifier les hallucinations contextuelles et guider le processus de génération fidèle. Diverses approches de post-édition ont été explorées pour améliorer la fidélité, telles qu'un flux de travail de recherche et de révision, où l'étape de recherche soulève des questions sur la réponse initiale du modèle et recueille des preuves pour chaque requête.

5.3.4 Cohérence logique

La **chaîne de pensée** (*chain of thought*) est une méthode qui encourage les LLMs à décomposer les problèmes complexes en étapes intermédiaires, améliorant ainsi la fiabilité du raisonnement (Wei *et al.*, 2022). Cependant, des recherches récentes ont montré que les raisonnements intermédiaires générés par les LLMs ne reflètent pas fidèlement leur comportement sous-jacent.

Pour améliorer la cohérence des raisonnements intermédiaires, Wang *et al.* (2024a) ont utilisé un cadre de distillation des connaissances entre un modèle-enseignant et un modèle-élève. Le modèle-enseignant génère un raisonnement cohérent à l'aide d'un décodage contrastif, tandis que le modèle-élève est affiné avec un objectif de raisonnement contrefactuel. Cette approche permet d'éliminer les raccourcis de raisonnement et de réduire les copies superficielles.

Li *et al.* (2024a) ont réalisé une analyse approfondie de la pertinence causale entre le contexte et la réponse pendant le raisonnement infidèle. Ils ont proposé un rapprochement inférentiel pour améliorer le raisonnement contextuel et filtrer les contextes bruyants dont la cohérence sémantique et les scores d'attribution au contexte sont faibles. De leur côté, Paul *et al.* (2024) ont décomposé le processus de raisonnement en deux modules : (1) un module d'inférence utilisant l'**optimisation des préférences directes** pour aligner le LLM vers des chaînes de raisonnement correctes sur des chaînes contrefactuelles et (2) un module de raisonnement encourageant le LLM à raisonner fidèlement sur des étapes de raisonnement en utilisant un objectif de préférence contrefactuelle et causale. Enfin, Xu *et al.* (2024a) incorporent des expressions symboliques dans les contextes pour décrire les étapes de raisonnement intermédiaires. Ils traduisent le contexte en langue naturelle en une représentation symbolique, formulent un plan par étapes pour résoudre le problème de raisonnement logique et vérifient la traduction et la chaîne de raisonnement pour garantir un raisonnement fidèle et logique.

6 Conclusion

Malgré les progrès rapides dans le domaine des LLMs et la publication continue de modèles puissants, les hallucinations restent un sujet pressant qui mérite une investigation continue. Notre étude a exploré

les origines et les mécanismes des hallucinations survenant dans les sorties des LLMs et les moyens de les atténuer. Les approches traditionnelles d'évaluation peuvent donner un aperçu des capacités des modèles dans des cas spécifiques, mais ne sont pas nécessairement utiles dans le cas des hallucinations. D'autre part, l'utilisation de LLM pour évaluer et atténuer leur propre sortie ou celle d'autres modèles peut être compromise. Bien que ces approches puissent offrir une amélioration significative avec un faible coût de calcul, elles dépendent fortement de la fidélité des LLM, qui restent toujours sujets à des erreurs. Notre conclusion, fondée sur la bibliographie, est que le développement de points de référence spécifiques aux LLMs, ainsi que de meilleures stratégies d'exploitation des connaissances à partir des données, est nécessaire.

Remerciements

Ce travail a été réalisé avec le soutien financier du projet FRAE-RAKEL (R093-FRAE-RAKEL) et ANITI (*Artificial and Natural Intelligence Toulouse Institute*), financé par le programme «Investissements d'avenir-PIA3» (ANR-19-PI3A-000). Nous tenons également à remercier le projet ANR COCOBOTS (ANR-21-FAI2-0005). Nos remerciements aussi à Bastien Navarri et aux relecteurs anonymes pour leur aide à la révision de la langue de l'article.

Références

- ABBAS A. K. M., TIRUMALA K., SIMIG D., GANGULI S. & MORCOS A. S. (2023). Semdedup : Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.
- AGRAWAL A., SUZGUN M., MACKAY L. & KALAI A. (2024). Do language models know when they're hallucinating references? In Y. GRAHAM & M. PURVER, Éd(s.), *Findings of the Association for Computational Linguistics : EACL 2024*, p. 912–928, St. Julian's, Malta : Association for Computational Linguistics.
- ALSHAHRANI M., THAFAR M. A. & ESSACK M. (2021). Application and evaluation of knowledge graph embeddings in biomedical data. *PeerJ Computer Science*, **7**, e341.
- ASHER N. & BHAR S. (2024). Strong hallucinations from negation and how to fix them. *arXiv preprint arXiv :2402.10543*.
- BAEK J., AJI A. F. & SAFFARI A. (2023). Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In B. DALVI MISHRA, G. DURRETT, P. JANSEN, D. NEVES RIBEIRO & J. WEI, Éd(s.), *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, p. 78–106, Toronto, Canada : Association for Computational Linguistics. DOI : [10.18653/v1/2023.nlrse-1.7](https://doi.org/10.18653/v1/2023.nlrse-1.7).
- BANG Y., CAHYAWIJAYA S., LEE N., DAI W., SU D., WILIE B., LOVENIA H., JI Z., YU T., CHUNG W., DO Q. V., XU Y. & FUNG P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In J. C. PARK, Y. ARASE, B. HU, W. LU, D. WIJAYA, A. PURWARIANTI & A. A. KRISNADHI, Éd(s.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 675–718, Nusa Dua, Bali : Association for Computational Linguistics. DOI : [10.18653/v1/2023.ijcnlp-main.45](https://doi.org/10.18653/v1/2023.ijcnlp-main.45).
- BOKSA P. (2009). On the neurobiology of hallucinations. *Journal of psychiatry & neuroscience : JPN*, **34**(4), 260.
- BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. ET AL. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.
- CHANG C.-C., REITTER D., AKSITOV R. & SUNG Y.-H. (2023). Kl-divergence guided temperature sampling. *arXiv preprint arXiv :2306.01286*.
- CHANG H., PARK J., YE S., YANG S., SEO Y., CHANG D.-S. & SEO M. (2024). How do large language models acquire factual knowledge during pretraining? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- 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*.
- CHIANG D. & CHOLAK P. (2022). Overcoming a theoretical limitation of self-attention. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éd(s.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 7654–7664, Dublin, Ireland : Association for Computational Linguistics. DOI : [10.18653/v1/2022.acl-long.527](https://doi.org/10.18653/v1/2022.acl-long.527).
- CHOI S., FANG T., WANG Z. & SONG Y. (2023). Kcts : Knowledge-constrained tree search decoding with token-level hallucination detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 14035–14053.

CHUANG Y.-S., XIE Y., LUO H., KIM Y., GLASS J. R. & HE P. (2024). Dola : Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

CHUNG H. W., HOU L., LONGPRE S., ZOPH B., TAY Y., FEDUS W., LI Y., WANG X., DEGHANI M., BRAHMA S. *ET AL.* (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, **25**(70), 1–53.

COHEN R., HAMRI M., GEVA M. & GLOBERSON A. (2023). LM vs LM : Detecting factual errors via cross examination. In H. BOUAMOR, J. PINO & K. BALI, ÉdS., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 12621–12640, Singapore : Association for Computational Linguistics. DOI : [10.18653/v1/2023.emnlp-main.778](https://doi.org/10.18653/v1/2023.emnlp-main.778).

DE CAO N., AZIZ W. & TITOV I. (2021). Editing Factual Knowledge in Language Models. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, ÉdS., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6491–6506, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : [10.18653/v1/2021.emnlp-main.522](https://doi.org/10.18653/v1/2021.emnlp-main.522).

DHULIAWALA S., KOMEILI M., XU J., RAILEANU R., LI X., CELIKYILMAZ A. & WESTON J. E. (2024). Chain-of-verification reduces hallucination in large language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

DUGAN L., IPPOLITO D., KIRUBARAJAN A., SHI S. & CALLISON-BURCH C. (2023). Real or fake text ? : Investigating human ability to detect boundaries between human-written and machine-generated text. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**(11), 12763–12771.

FENG Z., FENG X., ZHAO D., YANG M. & QIN B. (2024). Retrieval-generation synergy augmented large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 11661–11665 : IEEE.

FU J., NG S.-K., JIANG Z. & LIU P. (2024). GPTScore : Evaluate as you desire. 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. 6556–6576, Mexico City, Mexico : Association for Computational Linguistics. DOI : [10.18653/v1/2024.naacl-long.365](https://doi.org/10.18653/v1/2024.naacl-long.365).

GAO L., BIDERMAN S., BLACK S., GOLDING L., HOPPE T., FOSTER C., PHANG J., HE H., THITE A., NABESHIMA N. *ET AL.* (2020). The pile : An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv :2101.00027*.

GAO L., DAI Z., PASUPAT P., CHEN A., CHAGANTY A. T., FAN Y., ZHAO V., LAO N., LEE H., JUAN D.-C. & GUU K. (2023). RARR : Researching and revising what language models say, using language models. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, ÉdS., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 16477–16508, Toronto, Canada : Association for Computational Linguistics. DOI : [10.18653/v1/2023.acl-long.910](https://doi.org/10.18653/v1/2023.acl-long.910).

GEKHMAN Z., YONA G., AHARONI R., EYAL M., FEDER A., REICHAERT R. & HERZIG J. (2024). Does fine-tuning llms on new knowledge encourage hallucinations ? *arXiv preprint arXiv :2405.05904*.

GOODRICH B., RAO V., LIU P. J. & SALEH M. (2019). Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, p. 166–175, New York, NY, USA : Association for Computing Machinery. DOI : [10.1145/3292500.3330955](https://doi.org/10.1145/3292500.3330955).

GUERREIRO N. M., VOITA E. & MARTINS A. (2023). Looking for a needle in a haystack : A comprehensive study of hallucinations in neural machine translation. In A. VLACHOS & I. AUGENSTEIN, ÉdS., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational*

Linguistics, p. 1059–1075, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : [10.18653/v1/2023.eacl-main.75](https://doi.org/10.18653/v1/2023.eacl-main.75).

GUNASEKAR S., ZHANG Y., ANEJA J., MENDES C. C. T., DEL GIORNO A., GOPI S., JAVAHERIPI M., KAUFFMANN P., DE ROSA G., SAARIKIVI O. *ET AL.* (2023). Textbooks are all you need. *arXiv preprint arXiv :2306.11644*.

GUU K., LEE K., TUNG Z., PASUPAT P. & CHANG M. (2020). Retrieval augmented language model pre-training. In *International conference on machine learning*, p. 3929–3938 : PMLR.

GYAWALI B., ANASTASIOU L. & KNOTH P. (2020). Deduplication of scholarly documents using locality sensitive hashing and word embeddings. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éd.s., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 901–910, Marseille, France : European Language Resources Association.

HUANG L., YU W., MA W., ZHONG W., FENG Z., WANG H., CHEN Q., PENG W., FENG X., QIN B. *ET AL.* (2025). A survey on hallucination in large language models : Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, **43**(2), 1–55.

HUO S., ARABZADEH N. & CLARKE C. L. (2023). Retrieving supporting evidence for llms generated answers. *arXiv preprint arXiv :2306.13781*.

Ji Z., YU T., XU Y., LEE N., ISHII E. & FUNG P. (2023). Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 1827–1843.

JIANG C., QI B., HONG X., FU D., CHENG Y., MENG F., YU M., ZHOU B. & ZHOU J. (2024). On large language models' hallucination with regard to known facts. In K. DUH, H. GOMEZ & S. BETHARD, Éd.s., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 1041–1053, Mexico City, Mexico : Association for Computational Linguistics. DOI : [10.18653/v1/2024.naacl-long.60](https://doi.org/10.18653/v1/2024.naacl-long.60).

JIANG Z., XU F., GAO L., SUN Z., LIU Q., DWIVEDI-YU J., YANG Y., CALLAN J. & NEUBIG G. (2023). Active retrieval augmented generation. In H. BOUAMOR, J. PINO & K. BALI, Éd.s., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 7969–7992, Singapore : Association for Computational Linguistics. DOI : [10.18653/v1/2023.emnlp-main.495](https://doi.org/10.18653/v1/2023.emnlp-main.495).

KADAVATH S., CONERLY T., ASKELL A., HENIGHAN T., DRAIN D., PEREZ E., SCHIEFER N., HATFIELD-DODDS Z., DASARMA N., TRAN-JOHNSON E. *ET AL.* (2022). Language models (mostly) know what they know. *arXiv preprint arXiv :2207.05221*.

KANDPAL N., DENG H., ROBERTS A., WALLACE E. & RAFFEL C. (2023). Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, p. 15696–15707 : PMLR.

KANG H., BLEVINS T. & ZETTMLOYER L. (2024). Comparing hallucination detection metrics for multilingual generation. *arXiv preprint arXiv :2402.10496*.

LABAN P., KRYŚCIŃSKI W., AGARWAL D., FABBRI A. R., XIONG C., JOTY S. & WU C.-S. (2023). Llms as factual reasoners : Insights from existing benchmarks and beyond. *arXiv preprint arXiv :2305.14540*.

LADHAK F., DURMUS E., SUZGUN M., ZHANG T., JURAFSKY D., McKEOWN K. & HASHIMOTO T. (2023). When do pre-training biases propagate to downstream tasks ? a case study in text summarization. In A. VLACHOS & I. AUGENSTEIN, Éd.s., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 3206–3219, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : [10.18653/v1/2023.eacl-main.234](https://doi.org/10.18653/v1/2023.eacl-main.234).

LEE N., PING W., XU P., PATWARY M., FUNG P. N., SHOEYBI M. & CATANZARO B. (2022). Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, **35**, 34586–34599.

LI J., CAO P., CHEN Y., LIU K. & ZHAO J. (2024a). Towards faithful chain-of-thought : Large language models are bridging reasoners. *arXiv preprint arXiv :2405.18915*.

LI J., CHEN J., REN R., CHENG X., ZHAO X., NIE J.-Y. & WEN J.-R. (2024b). The dawn after the dark : An empirical study on factuality hallucination in large language models. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Édts., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 10879–10899, Bangkok, Thailand : Association for Computational Linguistics. DOI : [10.18653/v1/2024.acl-long.586](https://doi.org/10.18653/v1/2024.acl-long.586).

LI K., PATEL O., VIÉGAS F., PFISTER H. & WATTENBERG M. (2024c). Inference-time intervention : Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, **36**.

LIN S., HILTON J. & EVANS O. (2022). TruthfulQA : Measuring how models mimic human falsehoods. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Édts., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3214–3252, Dublin, Ireland : Association for Computational Linguistics. DOI : [10.18653/v1/2022.acl-long.229](https://doi.org/10.18653/v1/2022.acl-long.229).

LIU N. F., LIN K., HEWITT J., PARANJAPPE A., BEVILACQUA M., PETRONI F. & LIANG P. (2024). Lost in the middle : How language models use long contexts. *Transactions of the Association for Computational Linguistics*, **12**, 157–173. DOI : [10.1162/tacl_a_00638](https://doi.org/10.1162/tacl_a_00638).

LIU Y., YAO Y., TON J.-F., ZHANG X., CHENG R. G. H., KLOCHKOV Y., TAUFIQ M. F. & LI H. (2023). Trustworthy llms : A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv :2308.05374*.

LUO J., XIAO C. & MA F. (2024). Zero-resource hallucination prevention for large language models. In *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 3586–3602.

LUO Z., XIE Q. & ANANIADOU S. (2023). Chatgpt as a factual inconsistency evaluator for text summarization. *arXiv preprint arXiv :2303.15621*.

MANAKUL P., LIUSIE A. & GALES M. (2023). SelfCheckGPT : Zero-resource black-box hallucination detection for generative large language models. In H. BOUAMOR, J. PINO & K. BALI, Édts., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 9004–9017, Singapore : Association for Computational Linguistics. DOI : [10.18653/v1/2023.emnlp-main.557](https://doi.org/10.18653/v1/2023.emnlp-main.557).

MAYNEZ J., NARAYAN S., BOHNET B. & McDONALD R. (2020). On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv :2005.00661*.

McKENNA N., LI T., CHENG L., HOSSEINI M., JOHNSON M. & STEEDMAN M. (2023). Sources of hallucination by large language models on inference tasks. In H. BOUAMOR, J. PINO & K. BALI, Édts., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 2758–2774, Singapore : Association for Computational Linguistics. DOI : [10.18653/v1/2023.findings-emnlp.182](https://doi.org/10.18653/v1/2023.findings-emnlp.182).

MENG K., BAU D., ANDONIAN A. & BELINKOV Y. (2022a). Locating and editing factual associations in gpt. In S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO & A. OH, Édts., *Advances in Neural Information Processing Systems*, volume 35, p. 17359–17372 : Curran Associates, Inc.

MENG K., SHARMA A. S., ANDONIAN A., BELINKOV Y. & BAU D. (2022b). Mass-editing memory in a transformer. *arXiv preprint arXiv :2210.07229*.

MIN S., KRISHNA K., LYU X., LEWIS M., YIH W.-T., KOH P., IYYER M., ZETTLEMOYER L. & HAJISHIRZI H. (2023). FActScore : Fine-grained atomic evaluation of factual precision in long form text generation. In H. BOUAMOR, J. PINO & K. BALI, Édts., *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, p. 12076–12100, Singapore : Association for Computational Linguistics. DOI : [10.18653/v1/2023.emnlp-main.741](https://doi.org/10.18653/v1/2023.emnlp-main.741).

MISHRA A., PATEL D., VIJAYAKUMAR A., LI X. L., KAPANIPATHI P. & TALAMADUPULA K. (2021). Looking beyond sentence-level natural language inference for question answering and text summarization. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Édts., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1322–1336, Online : Association for Computational Linguistics. DOI : [10.18653/v1/2021.naacl-main.104](https://doi.org/10.18653/v1/2021.naacl-main.104).

MUHLGAY D., RAM O., MAGAR I., LEVINE Y., RATNER N., BELINKOV Y., ABEND O., LEYTON-BROWN K., SHASHUA A. & SHOHAM Y. (2024). Generating benchmarks for factuality evaluation of language models. In Y. GRAHAM & M. PURVER, Édts., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 49–66, St. Julian's, Malta : Association for Computational Linguistics.

NAIM O. & ASHER N. (2024). On explaining with attention matrices. In *ECAI 2024*, p. 1035–1042. IOS Press.

NAIM O. & ASHER N. (2025). Two in context learning tasks with complex functions. *arXiv preprint arXiv :2502.03503*.

NAIM O., FOUILHÉ G. & ASHER N. (2024). Re-examining learning linear functions in context. *arXiv preprint arXiv :2411.11465*.

NAN F., NALLAPATI R., WANG Z., DOS SANTOS C., ZHU H., ZHANG D., MCKEOWN K. & XIANG B. (2021). Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 2727–2733.

ONOE Y., ZHANG M., CHOI E. & DURRETT G. (2022). Entity cloze by date : What LMs know about unseen entities. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Édts., *Findings of the Association for Computational Linguistics : NAACL 2022*, p. 693–702, Seattle, United States : Association for Computational Linguistics. DOI : [10.18653/v1/2022.findings-naacl.52](https://doi.org/10.18653/v1/2022.findings-naacl.52).

PAN S., LUO L., WANG Y., CHEN C., WANG J. & WU X. (2024). Unifying large language models and knowledge graphs : A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

PANICKSERY N., GABRIELI N., SCHULZ J., TONG M., HUBINGER E. & TURNER A. M. (2023). Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv :2312.06681*.

PAUL D., WEST R., BOSSELUT A. & FALTINGS B. (2024). Making reasoning matter : Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 15012–15032.

PEREZ E., RINGER S., LUKOŠIŪTĖ K., NGUYEN K., CHEN E., HEINER S., PETTIT C., OLSSON C., KUNDU S., KADAVATH S. ET AL. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv :2212.09251*.

PRESS O., ZHANG M., MIN S., SCHMIDT L., SMITH N. & LEWIS M. (2023). Measuring and narrowing the compositionality gap in language models. In H. BOUAMOR, J. PINO & K. BALI, Édts., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 5687–5711, Singapore : Association for Computational Linguistics. DOI : [10.18653/v1/2023.findings-emnlp.378](https://doi.org/10.18653/v1/2023.findings-emnlp.378).

QIAO S., OU Y., ZHANG N., CHEN X., YAO Y., DENG S., TAN C., HUANG F. & CHEN H. (2022). Reasoning with language model prompting : A survey. *arXiv preprint arXiv :2212.09597*.

- RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, **21**(140), 1–67.
- RAHIMI Z., AMIRZADEH H., SOHRABI A., TAGHAVI Z. & SAMETI H. (2024). HalluSafe at SemEval-2024 task 6 : An NLI-based approach to make LLMs safer by better detecting hallucinations and overgeneration mistakes. In A. K. OJHA, A. S. DOĞRUÖZ, H. TAYYAR MADABUSHI, G. DA SAN MARTINO, S. ROSENTHAL & A. ROSÁ, ÉdS., *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, p. 139–147, Mexico City, Mexico : Association for Computational Linguistics. DOI : [10.18653/v1/2024.semeval-1.22](https://doi.org/10.18653/v1/2024.semeval-1.22).
- RAM O., LEVINE Y., DALMEDIGOS I., MUHLGAY D., SHASHUA A., LEYTON-BROWN K. & SHOHAM Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, **11**, 1316–1331. DOI : [10.1162/tacl_a_00605](https://doi.org/10.1162/tacl_a_00605).
- RAWTE V., SHETH A. & DAS A. (2023). A survey of hallucination in large foundation models. *arXiv preprint arXiv :2309.05922*.
- ROSSI A., BARBOSA D., FIRMANI D., MATINATA A. & MERIALDO P. (2021). Knowledge graph embedding for link prediction : A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **15**(2), 1–49.
- SAUNDERS W., YEH C., WU J., BILLS S., OUYANG L., WARD J. & LEIKE J. (2022). Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv :2206.05802*.
- SHARMA M., TONG M., KORBAK T., DUVENAUD D., ASKELL A., BOWMAN S. R., DURMUS E., HATFIELD-DODDS Z., JOHNSTON S. R., KRAVEC S. M. ET AL. (2024). Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.
- SHI W., MIN S., LOMELI M., ZHOU C., LI M., SZILVASY G., JAMES R., LIN X. V., SMITH N. A., ZETTMLOYER L. ET AL. (2023). In-context pretraining : Language modeling beyond document boundaries. *arXiv preprint arXiv :2310.10638*.
- TAN C., ZHANG G. & FU J. (2023). Massive editing for large language models via meta learning. *arXiv preprint arXiv :2311.04661*.
- TONMOY S., ZAMAN S., JAIN V., RANI A., RAWTE V., CHADHA A. & DAS A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv :2401.01313*.
- TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. ET AL. (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.
- VARSHNEY N., YAO W., ZHANG H., CHEN J. & YU D. (2023). A stitch in time saves nine : Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv :2307.03987*.
- VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.
- WANG S., ZHU Y., LIU H., ZHENG Z., CHEN C. & LI J. (2024a). Knowledge editing for large language models : A survey. *ACM Computing Surveys*, **57**(3), 1–37.
- WANG Y., GANGI REDDY R., MUJAHID Z. M., ARORA A., RUBASHEVSKII A., GENG J., MOHAMMED AFZAL O., PAN L., BORENSTEIN N., PILLAI A., AUGENSTEIN I., GUREVYCH I. & NAKOV P. (2024b). Factcheck-bench : Fine-grained evaluation benchmark for automatic fact-checkers. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, ÉdS., *Findings of the Association for Computational*

Linguistics : EMNLP 2024, p. 14199–14230, Miami, Florida, USA : Association for Computational Linguistics. DOI : [10.18653/v1/2024.findings-emnlp.830](https://doi.org/10.18653/v1/2024.findings-emnlp.830).

WANG Y., WANG M., ARSLAN MANZOOR M., GEORGIEV G., JYOTI DAS R. & NAKOV P. (2024c). Factuality of large language models in the year 2024. *arXiv e-prints*, p. arXiv–2402.

WEI J., HUANG D., LU Y., ZHOU D. & LE Q. V. (2023). Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv :2308.03958*.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E., LE Q. V. & ZHOU D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO & A. OH, Éd.s., *Advances in Neural Information Processing Systems*, volume 35, p. 24824–24837 : Curran Associates, Inc.

WEN Y., WANG Z. & SUN J. (2024). MindMap : Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éd.s., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 10370–10388, Bangkok, Thailand : Association for Computational Linguistics. DOI : [10.18653/v1/2024.acl-long.558](https://doi.org/10.18653/v1/2024.acl-long.558).

XIAO Y. & WANG W. Y. (2021). On hallucination and predictive uncertainty in conditional language generation. In P. MERLO, J. TIEDEMANN & R. TSARFATY, Éd.s., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 2734–2744, Online : Association for Computational Linguistics. DOI : [10.18653/v1/2021.eacl-main.236](https://doi.org/10.18653/v1/2021.eacl-main.236).

XIONG M., HU Z., LU X., LI Y., FU J., HE J. & HOOI B. (2023). Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv :2306.13063*.

XU J., FEI H., PAN L., LIU Q., LEE M.-L. & HSU W. (2024a). Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 13326–13365.

XU Z., JAIN S. & KANKANHALLI M. (2024b). Hallucination is inevitable : An innate limitation of large language models. *arXiv preprint arXiv :2401.11817*.

YAO J.-Y., NING K.-P., LIU Z.-H., NING M.-N., LIU Y.-Y. & YUAN L. (2023). Llm lies : Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv :2310.01469*.

YU F., ZHANG H., TIWARI P. & WANG B. (2024). Natural language reasoning, a survey. *ACM Computing Surveys*, **56**(12), 1–39.

YU W., ZHANG Z., LIANG Z., JIANG M. & SABHARWAL A. (2023). Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv :2305.14002*.

ZHANG H., DIAO S., LIN Y., FUNG Y. R., LIAN Q., WANG X., CHEN Y., JI H. & ZHANG T. (2023a). R-tuning : Teaching large language models to refuse unknown questions. *arXiv preprint arXiv :2311.09677*.

ZHANG H., DUCKWORTH D., IPPOLITO D. & NEELAKANTAN A. (2021). Trading off diversity and quality in natural language generation. *EACL 2021*, p.25.

ZHANG Q., SINGH C., LIU L., LIU X., YU B., GAO J. & ZHAO T. (2023b). Tell your model where to attend : Post-hoc attention steering for llms. *arXiv preprint arXiv :2311.02262*.

ZHANG T., LADHAK F., DURMUS E., LIANG P., MCKEOWN K. & HASHIMOTO T. B. (2024a). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, **12**, 39–57.

ZHANG T., QIU L., GUO Q., DENG C., ZHANG Y., ZHANG Z., ZHOU C., WANG X. & FU L. (2023c). Enhancing uncertainty-based hallucination detection with stronger focus. In H. BOUAMOR, J. PINO & K. BALI, Éd.s., *Proceedings of the 2023 Conference on Empirical Methods in Natural*

Language Processing, p. 915–932, Singapore : Association for Computational Linguistics. DOI : [10.18653/v1/2023.emnlp-main.58](https://doi.org/10.18653/v1/2023.emnlp-main.58).

ZHANG Y., LI Y., CUI L., CAI D., LIU L., FU T., HUANG X., ZHAO E., ZHANG Y., CHEN Y. *ET AL.* (2023d). Siren’s song in the ai ocean : a survey on hallucination in large language models. *arXiv preprint arXiv :2309.01219*.

ZHANG Y., WANG X., LIANG J., XIA S., CHEN L. & XIAO Y. (2024b). Chain-of-knowledge : Integrating knowledge reasoning into large language models by learning from knowledge graphs. *arXiv preprint arXiv :2407.00653*.

ZHAO R., LI X., JOTY S., QIN C. & BING L. (2023). Verify-and-edit : A knowledge-enhanced chain-of-thought framework. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Édts., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 5823–5840, Toronto, Canada : Association for Computational Linguistics. DOI : [10.18653/v1/2023.acl-long.320](https://doi.org/10.18653/v1/2023.acl-long.320).

ZHAO Y., ZHANG J., CHERN I., GAO S., LIU P., HE J. *ET AL.* (2024). Felm : Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, **36**.

ZHOU C., NEUBIG G., GU J., DIAB M., GUZMÁN F., ZETTLEMOYER L. & GHAZVININEJAD M. (2021). Detecting hallucinated content in conditional neural sequence generation. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Édts., *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 1393–1404, Online : Association for Computational Linguistics. DOI : [10.18653/v1/2021.findings-acl.120](https://doi.org/10.18653/v1/2021.findings-acl.120).

A Exemples d'hallucinations

Cette section présente des exemples des différents types d'hallucinations tels qu'établis dans la Section 2. Les instructions ont été fournies en anglais. Tous les modèles ont été testés en février 2025.

A.1 Hallucinations de fidélité

Utilisateur :	Translate everything in French. What is the height of the Eiffel tower?
Modèle :	The height of the Eiffel Tower is 324 meters. In French, the sentence would be : "Quelle est la hauteur de la tour Eiffel ?" And the height of the Eiffel Tower in French would be : "La hauteur de la tour Eiffel est de 324 mètres."

TABLE 1 – Exemple d'hallucination avec **incohérence d'instruction**. Le modèle ne traduit pas l'intégralité du message et répond en même temps à la question, puis le traduit comme l'utilisateur l'a demandé.

Modèle : Mistral-Small-24B-Instruct-2501. Source : huggingface.co

Utilisateur :	I only have vegetables in my fridge. Give me a recipe for a meat lover's pizza.
Modèle :	A meat lover's pizza traditionally features a combination of various meats like pepperoni, sausage, bacon, and ham. Since you only have vegetables, we can create a veggie-inspired version that mimics the hearty, satisfying vibe of a meat lover's pizza. Here's a Veggie Meat Lover's Pizza recipe : Ingredients : Pizza dough (store-bought or homemade) 1/2 cup pizza sauce or marinara sauce 1 cup shredded mozzarella cheese 1/2 cup shredded cheddar cheese (optional, for a richer flavor) 1/4 cup red onion, thinly sliced 1/2 cup bell peppers (green, red, or yellow), thinly sliced [...]

TABLE 2 – Exemple d'hallucination avec **incohérence de contexte**. L'utilisateur déclare qu'il n'a que des légumes dans son réfrigérateur, ce qui implique que ce sont les seuls ingrédients disponibles pour la recette. Le modèle produit une recette qui nécessite de la pâte et du fromage. Réponse abrégée par souci de concision.

Modèle : GPT-4o mini. Source : chatgpt.com

Utilisateur : Ken created a package to send to his brother to help him repair his house. Ken placed a box on a scale, and then he poured into the box enough nails to bring the weight to 2 pounds. Then, he added enough screws and bolts to cause the weight to triple. Next, he added to the package a set of socket wrenches whose total weight equaled $1/2$ of the weight of the package so far. And finally, he added a bunch of drills and bits to double the weight once again. What is the weight of the package ?

Modèle : Step 1 : Determine the initial weight of the nails.
The initial weight of the nails is 2 pounds.
Step 2 : Calculate the weight after adding screws and bolts.
After adding screws and bolts, the weight triples. So, the new weight is $2 * 3 = 6$ pounds.
Step 3 : Calculate the weight after adding the socket wrenches.
The weight of the socket wrenches is $1/2$ of the current weight. So, the weight of the socket wrenches is $6 / 2 = 3$ pounds. The total weight now is $6 + 3 = 9$ pounds.
Step 4 : Calculate the weight after adding the drills and bits.
The weight doubles after adding the drills and bits. So, the new weight is $9 * 2 = 18$ pounds.
However, we know the final weight is 42 pounds, not 18 pounds, which means our calculation approach needs adjustment to find the weight of the drills and bits.
Step 5 : Correct the approach to find the weight of the drills and bits.
Given that the final weight of the package is 42 pounds and this is after the weight has doubled from the addition of the drills and bits, we can work backwards. If the weight doubles to 42 pounds, then before adding the drills and bits, the weight was $42 / 2 = 21$ pounds.
Step 6 : Calculate the weight of the drills and bits.
Since the weight before adding the drills and bits was 21 pounds, and after adding them it became 42 pounds, the weight of the drills and bits is $42 - 21 = 21$ pounds.
The final answer is : 21

TABLE 3 – Un exemple d'hallucination présentant une **incohérence logique**. Le modèle peut calculer correctement le poids final de la boîte, en suivant les instructions de l'utilisateur et en calculant les additions étape par étape. Cependant, après avoir obtenu la bonne réponse, il hallucine une réponse différente et commence à calculer sur la base de la réponse incorrecte. Réponse abrégée par souci de concision.

Modèle : Meta Llama-3 70B. Source : huggingface.co

A.2 Erreurs de faits

Utilisateur :	Who is a famous Slovenian scientist ?
Modèle :	One famous Slovenian scientist is Janez Vajkard Valvasor (1641-1693), a polymath and scholar who made significant contributions to various fields, including natural history, geography, and ethnology. However, a more modern example is : Janez Puhar (1814-1864), a Slovenian priest, photographer, and inventor, who is best known for inventing a photographic process called "heliography" or "Puhar's process", which allowed for the creation of photographic images on glass plates. Another notable example is : Milko Škofič (1905-1981), a Slovenian geographer and climatologist, who made significant contributions to the field of geography, particularly in the study of climate and weather patterns. But perhaps one of the most famous Slovenian scientists is : Josip Plemelj (1873-1967), a Slovenian mathematician, who made significant contributions to the field of mathematics, particularly in the areas of differential equations, potential theory, and topology. More recently, a notable example is : Alojzij Šuštar (1948-), a Slovenian chemist and materials scientist, who has made significant contributions to the field of materials science, particularly in the development of new materials and technologies. These are just a few examples of the many talented Slovenian scientists who have made significant contributions to their respective fields.

TABLE 4 – Exemple d’hallucination avec des **erreurs d’entité**. Lorsque le modèle est interrogé sur les scientifiques slovènes, certaines de ses réponses nomment correctement des scientifiques slovènes existants, mais d’autres font référence à des Slovènes célèbres de la vie réelle qui ne sont pas des scientifiques. En outre, le LLM attribue des dates de naissance/décès erronées aux non-scientifiques et leur imagine une profession et des réalisations.

Modèle : Llama-3.3-70B-Instruct. Source : huggingface.co

Utilisateur :	When did Léon Marchand win an Olympic medal ?
Modèle :	Léon Marchand won an Olympic medal at the 2020 Summer Olympics in Tokyo, Japan. He won a silver medal in the 400m individual medley event on July 25, 2021, with a time of 4 :10.69.

TABLE 5 – Exemple d’hallucination avec des **erreurs de relation**. Le médaillé d’argent de cette épreuve n’est pas Léon Marchand, mais il a participé à l’épreuve et s’est classé sixième.

Modèle : Llama-3.3-70B-Instruct. Source : huggingface.co

Utilisateur :	Give me the 5 most important historical landmarks of Athens, with dates of creation and visitor information.
Modèle :	[...] Here's a concise summary of the 5 most important historical landmarks in Athens, with dates and key visitor details : Acropolis (5th century BCE) Visitor Info : Open daily (summer : 8 AM–8 PM; winter : 8 AM–5 PM). Tickets : €20 (summer), €10 (winter). Tips : Arrive early; wear sturdy shoes. [...] Panathenaic Stadium (330 BCE, rebuilt 144 CE, restored 1896) Visitor Info : Summer 8 AM–7 PM; winter 8 AM–5 PM. Tickets : €10 (includes audio guide). Tips : Walk the Olympic track. Kerameikos (5th century BCE) Visitor Info : Summer 8 AM–8 PM; winter 8 AM–3 PM. Tickets : €8 or combined pass. Tips : Explore funerary artifacts in the museum. [...]

TABLE 6 – Exemple d'hallucination avec une **erreur d'incomplétude**. Le modèle récupère correctement les informations sur les sites touristiques d'Athènes, les dates, les informations, les différents prix des billets, sauf pour le site archéologique de Kerameikos, où il ne mentionne pas tous les prix des billets. Réponse abrégée par souci de concision.

Modèle : DeepSeek-R1. Source : huggingface.co

Utilisateur :	Who is the prime minister of France?
Modèle :	The Prime Minister of France is Élisabeth Borne . She has been in office since May 16, 2022.

TABLE 7 – Exemple d'hallucination avec une **erreur d'obsolescence**. Le modèle répond à la question de qui est l'actuel premier ministre de la France en citant une ancienne première ministre de la France.

Modèle : Llama-3.3-70B-Instruct. Source : huggingface.co

A.3 Fabrication de faits

Utilisateur :	Where is the Parthenon?
Modèle :	The Parthenon is located in Ancient France . It stands at an elevation of 66 meters (215 feet) and was constructed from the 10th to the 14th century under the rule of Charlemagne the Charpec . The foundation of the Parthenon was laid by the French military during the Battle of Ess by using fireprongs to blow the gravel-filled mausoleum out of gravity . The Great Exterior Prism , designed by Louis XV , [...]

TABLE 8 – Exemple d'hallucination avec des **hallucinations invérifiables**. Le modèle hallucine l'emplacement du Parthénon (France antique), la hauteur du monument, sa période de construction, sa fonction et ses composants. Elle fabrique un souverain et une bataille inexistantes et ajoute un personnage historique réel qui n'appartient pas à l'époque qu'elle hallucine. Réponse abrégée par souci de concision.

Modèle : DeepSeek-R1-Distill-Qwen-1.5B. Source : huggingface.co

Utilisateur :	What is Thomas Edison's contribution to electric vehicles ?
Modèle :	<p>Okay, so I need to figure out Thomas Edison's contributions to electric vehicles. Let me start by recalling what I know about Edison. He was a prolific inventor, known for the light bulb, phonograph, and the motion picture camera. But electric vehicles ? Hmm, I remember that he worked on batteries. Oh right, the nickel-iron battery ! Wait, how does that relate to electric cars ?</p> <p>I think Edison was interested in improving batteries because early electric cars used lead-acid batteries, which were heavy and had limited range. Maybe he wanted to create a better battery to make electric vehicles more practical. But when did he do this ? I think it was around the early 1900s. Let me check... <u>Yes, he patented the nickel-iron battery in 1901. So his contribution was developing a better battery technology for electric vehicles. [...]</u></p>
Utilisateur :	Can you summarize your answer ?
Modèle :	<p>Okay, the user asked me to summarize my previous answer about Thomas Edison's contributions to electric vehicles. Let me start by recalling what I wrote earlier. I discussed Edison's development of the nickel-iron battery, his collaboration with Henry Ford, his promotion of electrification, and the legacy and limitations of his work. [...] Summary :</p> <p>Thomas Edison's key contribution to electric vehicles (EVs) was his invention of the nickel-iron battery (1901), designed to replace lead-acid batteries. It offered superior durability, sustainability, and longevity, aligning with his vision of EVs surpassing gasoline cars. Though his battery had lower energy density and higher costs, limiting its adoption, Edison collaborated with Henry Ford to explore affordable EVs. While market shifts toward gasoline engines stalled these efforts, his work laid groundwork for future battery tech and renewable energy storage. Edison's advocacy for electrification and clean transportation highlighted EVs' potential, influencing later innovations in sustainable mobility.</p>

TABLE 9 – Exemple d'hallucination avec des **hallucinations de surproclamation**. Edison n'a pas inventé la batterie nickel-fer, il a contribué à son développement et l'a fait breveter aux États-Unis en premier. Le modèle l'indique correctement dans la première réponse, mais de manière incorrecte lorsqu'on lui demande de produire un résumé. Réponses abrégées par souci de concision.

Modèle : DeepSeek-R1. Source : huggingface.co