Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Robustesse des LLM dans les contextes longs, hallucinations et détection sur questions-réponses séquentielles

Aygalic Jara--Mikolajczak[1,2] Thomas Lavergne[1]
Christophe Servan[1,3] Sophie Rosset[1]
(1) Université Paris-Saclay, CNRS, LISN, 91405, Orsay, France
(2) SCIAM, 75008, Paris, France
(3) AMIAD, Pôle Recherche, 91120, Palaiseau, France
firstname.lastname@lisn.fr, firstname.lastname@polytechnique.edu

RÉSUMÉ

Dans ce travail, nous évaluons l'impact de l'allongement du contexte sur le taux d'hallucination et les performances de détection d'hallucinations à travers sept modèles à poids ouverts sur le jeu de données TriviaQA. Nous utilisons des sondes linéaires, une approche efficace pour détecter les hallucinations dans les LLM, mais leur robustesse dans les contextes conversationnels longs reste peu étudiée. Nos résultats montrent que les taux d'hallucination restent stables jusqu'à 400 tours et 75 000 tokens, et que des sondes entraînées sur un tour unique généralisent bien aux contextes multi-tours. En revanche, l'injection de réponses oracles dans le contexte dégrade systématiquement les performances, suggérant que des réponses hors distribution perturbent les représentations internes du modèle.

ABSTRACT

LLM Robustness under Extended Contexts : Hallucination and Detection in Sequential Question Answering

In this work, we evaluate the impact of context length on hallucination rates and hallucination detection performance across seven open-weight models on the TriviaQA dataset. We use linear probes, which are an effective approach for detecting hallucinations in LLMs, but their robustness in long conversational contexts remains understudied. Our results show that hallucination rates remain stable up to 400 turns and 75,000 tokens, and that probes trained on single turns generalize well to multi-turn contexts. However, injecting oracle responses into the context systematically degrades performance, suggesting that out-of-distribution responses disrupt the model's internal representations.

MOTS-CLÉS : grands modèles de langue, hallucination, détection d'hallucinations, contexte long, questions-réponses séquentielles, sondes linéaires.

KEYWORDS: large language models, hallucination, hallucination detection, long-context, sequential question answering, linear probes.

1 Introduction

Les grands modèles de langue (ou LLM pour Large Language Models) sont des outils très polyvalents qui ont su gagner l'intérêt grâce à leur capacité à répondre à des questions de la vie de tous les

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 512–523.

jours (Open AI, 2022; Anthropic, 2023). Toutefois, malgré leurs performances impressionnantes, ces modèles sont sujets aux *hallucinations*. Un phénomène où ils génèrent des informations incorrectes ou infondées avec une grande assurance, entraînant ainsi l'utilisateur dans l'erreur.

Les hallucinations sont présentées comme l'une des principales limitations des LLM (Ji *et al.*, 2023) et un frein majeur à leur adoption. Elles se divisent en deux types : les hallucinations de fidélité et de factualité (Huang *et al.*, 2024). Dans cette étude, nous nous intéressons aux hallucinations de factualité dans le cadre de questions-réponses séquentielles sans contexte additionnel. Prévenir les utilisateurs lorsque ces hallucinations surviennent constitue donc un enjeu pratique important, et la détection des hallucinations fait à ce titre l'objet de nombreux travaux. Les méthodes proposées varient selon les contraintes d'accès au modèle : le sampling pour les systèmes boîte noire, les méthodes fondées sur l'entropie lorsque les distributions de tokens sont accessibles, et les sondes linéaires ou auto-encodeurs pour les modèles boîte blanche. Néanmoins, ces méthodes n'ont que rarement fait l'objet d'une étude approfondie dans le cadre de contextes longs (Liu *et al.*, 2025) ou de conversations multi-tours.

Dans ce travail, nous nous intéressons aux sondes linéaires (Alain & Bengio, 2018), qui constituent une approche ajoutant un coût minimal tout en ayant des performances de détection élevées pour identifier les hallucinations dans les LLM. Ces résultats ont notamment pu être vérifiés dans des contextes simples de questions-réponses à tour unique (Marks & Tegmark, 2024). Nous présentons une investigation systématique de l'influence de la taille du contexte conversationnel sur le comportement des modèles de langue. Elle s'articule autour de deux axes complémentaires. Tout d'abord, nous examinons la propension des modèles à générer des hallucinations au fur et à mesure que le contexte s'étend ; nous cherchons également à déterminer si la correction explicite d'erreurs dans l'historique de la conversation (par exemple, la réinjection de réponses rectifiées) modère cette propension. Ensuite, nous évaluons l'impact de l'accroissement du contexte sur les performances des sondes.

Nous nous situons dans un cadre de question-réponse conversationnel. On formule les hypothèses suivantes : **(H1)** la surcharge du contexte entraîne une augmentation des hallucinations, **(H2)** les sondes perdent leur capacité à discerner les hallucinations à mesure que le contexte croit et enfin **(H3)** la présence d'erreurs dans le contexte impacte (H1) et (H2). Afin de valider ces hypothèses, nous sélectionnons des modèles à poids ouverts récents couvrant diverses architectures et échelles, en privilégiant des tailles computationnellement accessibles, ainsi que le jeu de données TriviaQA comme détaillé dans la section 3.

Ce travail apporte quatre contributions principales. Nous proposons d'abord un protocole expérimental fondé sur TriviaQA pour évaluer le comportement des modèles et des sondes linéaires sur des contextes multi-tours. Ce protocole nous permet d'apporter des éléments de réponse à trois questions de recherche : **(1)** les hallucinations sont-elles plus fréquentes à mesure que le contexte conversationnel s'allonge ? **(2)** Les sondes de détection sont-elles robustes à cette augmentation ? **(3)** Remplacer les réponses du modèle par les réponses attendues atténue-t-il ces effets ?

Dans la section suivante, nous présentons les travaux portant sur les comportements hallucinatoires des modèles en fonction de la taille du contexte. Nous abordons également les méthodes de détection des hallucinations et leur étude dans les contextes de grande taille. Dans la section 3, nous détaillons la méthode développée pour l'élaboration des contextes, l'évaluation des modèles et des sondes. Les résultats obtenus font l'objet d'une présentation et d'une discussion approfondie dans cette même section. Enfin, nous présentons les limites (section 5) et la conclusion (section 6) de cette étude.

2 État de l'art

La gestion des contextes longs constitue un défi majeur pour les LLM. De nombreuses solutions ont été proposées, notamment au niveau architectural, comme les mécanismes d'attention parcimonieuse (Beltagy *et al.*, 2020) ou les méthodes d'attention linéaires (Team *et al.*, 2025b), avec des résultats encourageants.

Des travaux récents ont mis en évidence que les performances des LLM se dégradent significativement à mesure que la taille du contexte augmente. Notamment, Hsieh *et al.* (2024) montrent, à travers le benchmark RULER, que malgré des performances quasi-parfaites sur des tâches simples de récupération d'information (*needle-in-a-haystack*) (Kamradt, 2023), la quasi-totalité des modèles échoue à maintenir leurs performances sur des tâches plus complexes dès que le contexte dépasse 32K tokens, y compris pour des modèles affichant des fenêtres de contexte bien supérieures. Ce résultat suggère que la taille de contexte affichée par un modèle ne reflète pas nécessairement sa capacité effective à exploiter l'information sur l'ensemble de cette fenêtre. Toutefois, ces travaux s'intéressent principalement à la capacité de récupération et d'agrégation d'information dans des contextes longs synthétiques, et non à la détection des hallucinations factuelles dans un cadre conversationnel. À notre connaissance, la robustesse des sondes linéaires face à l'allongement du contexte multi-tours reste une question ouverte.

Il existe de nombreuses approches pour détecter ces hallucinations (Huang *et al.*, 2024), parmi lesquelles on distingue principalement deux groupes : les approches boîtes noires et boîtes blanches. Les approches boîte noire utilisent des méthodes de "self-consistency" (Dhuliawala *et al.*, 2023), de sampling (Manakul *et al.*, 2023), de fact checking (Min *et al.*, 2023) ou encore d'autres variantes basées sur l'entropie sémantique ou la similarité des phrases (Kossen *et al.*, 2024; Dale *et al.*, 2023). Elles ont l'avantage d'être utilisables sur des modèles exposés par API, mais l'inconvénient d'être soit plus coûteuses, soit moins performantes que les alternatives en boîte blanche. Les méthodes de détection des hallucinations en boîte blanche permettent principalement d'utiliser les signaux en interne pour quantifier l'incertitude du modèle, donnant ainsi un signal sur les hallucinations, parmi lesquelles on trouve l'utilisation d'auto-encodeurs parcimonieux (Ferrando *et al.*, 2024), l'utilisation de la matrice de covariance de l'espace latent (Chen *et al.*, 2023) ou encore les sondes linéaires (Alain & Bengio, 2018), qui ont prouvé leur efficacité dans la détection des hallucinations (Azaria & Mitchell, 2023; Binkowski *et al.*, 2025).

La question des contextes multi-tours a par ailleurs été abordée à travers des jeux de données conversationnels tels que CoQA (Reddy *et al.*, 2019) et QuAC (Choi *et al.*, 2018), dans lesquels chaque conversation porte sur un document unique. Ces ressources restent néanmoins limitées par la taille des contextes qu'elles impliquent (15 tours de conversation en moyenne pour CoQA pour 8K conversations contre 7 en moyenne pour QuAC pour 14K conversations), insuffisantes pour observer une éventuelle perturbation du comportement des modèles.

Il reste cependant un flou dans la littérature concernant la détection des hallucinations dans le cas des longs contextes (Liu *et al.*, 2025). Certaines approches hiérarchiques ont été utilisées dans le cadre du résumé (Liu *et al.*, 2025), mais encore peu d'autres cas d'usage ont été documentés. FACTScore (Min *et al.*, 2023) propose un cadre de détection des hallucinations en long contexte, en utilisant un système de découpage de la génération en propositions atomiques suivi d'une phase de vérification des propositions individuelles, entraînant ainsi un coût supplémentaire par rapport aux sondes linéaires non négligeable, mais qui a l'intérêt de fonctionner en scénario boîte noire. Il existe aussi des études

dans les systèmes de génération augmentée par récupération (ou RAG pour Retrieval-Augmented Generation) (Shi *et al.*, 2023). Il n'y a à notre connaissance que peu de réponses sur le comportement des LLMs à mesure que le contexte s'allonge, mettant ainsi en question la fiabilité des systèmes de questions-réponses et la robustesse de la détection d'hallucinations sur ces systèmes.

3 Méthode

Cette section décrit le protocole expérimental mis en place pour répondre aux hypothèses **H1**, **H2** et **H3**. L'objectif est d'évaluer, dans un cadre de question-réponse séquentielle, comment l'allongement du contexte affecte, d'une part, le taux d'hallucinations du modèle et, d'autre part, la capacité des sondes linéaires à les détecter. Pour ce faire, nous construisons des contextes multi-tours de longueur croissante à partir du jeu de données TriviaQA, que nous soumettons à plusieurs modèles à poids ouverts de tailles et d'architectures variées. Les sondes sont entraînées sur des contextes à tour unique, puis évaluées sur l'ensemble des tours, afin de tester leur capacité de généralisation. Nous comparons également deux stratégies de construction des contextes — l'une utilisant les réponses réelles du modèle, l'autre substituant les réponses correctes — pour mesurer l'effet de la propagation des erreurs.

3.1 Données

Pour construire nos contextes multi-tours, nous nous appuyons sur le jeu de données TriviaQA (Joshi *et al.*, 2017), composé de 174K triplets $\langle Document, Question, Rponse(s)\rangle$ extraits du *split train* disponible sur HuggingFace. Les documents sont des extraits de Wikipédia, auxquels sont associés des questions de culture générale ainsi qu'un ensemble d'alias constituant les réponses acceptées. Voici un exemple traduit : Question : "Quel est le premier vrai prénom de Bruce Willis ?" Réponse : "walter". Si le modèle répond "wallas" ou "bruce", ce qui est faux, cela sera comptabilisé comme une hallucination.

L'étude des approches avec contexte sourcé (type RAG (Lewis *et al.*, 2021)) étant largement documentée dans la littérature (Gao *et al.*, 2024), nous ne considérons pas ici l'apport de documents externes. Nous nous concentrons ainsi sur les paires $\langle Question, Reponse(s)\rangle$ afin d'évaluer la capacité des modèles à mobiliser leurs connaissances, indépendamment des informations présentes dans le contexte fourni.

Les données sont partagées de la manière suivante : 60% pour entraîner les sondes, 20% pour le jeu de test et 20% pour le jeu de validation. Les données sont ensuite mélangées au sein de chaque partition.

3.2 Construction des contextes multi-tours et stratégie oracle

La procédure pour construire les données des contextes multi-tours est la suivante : On considère un modèle de langue M et un contexte multi-tours défini par :

$$C_n = \{(q_1, r_1), (q_2, r_2), \ldots, (q_{n-1}, r_{n-1}), q_n\}$$

où q_i représente la question au tour i et $r_i = M(C_i)$ la réponse générée par le modèle. On étiquette chaque réponse r_n comme correcte ($y_n = 1$) ou hallucinée ($y_n = 0$). On précise que les données sont passées dans les outils de formatage de prompt propres à chaque modèle proposé par HuggingFace.

3.2.1 Stratégie de génération Oracle et Naturelle

Deux stratégies sont comparées pour construire C_n avec $n > 1$:

— **stratégie Naturelle** : les réponses du modèle sont utilisées telles quelles, qu'elles soient correctes ou non. Le contexte est construit de manière récursive : chaque nouvelle question est posée avec l'historique complet des échanges précédents. Formellement, on définit C_k^{nat} par récurrence :

$$C_1^{nat} = \big((q_1, M(q_1))\big)$$

$$C_k^{nat} = \big(C_{k-1}^{nat},\ (q_k, M(C_{k-1}^{nat}, q_k))\big) \quad \text{pour } k \geq 2$$

où $M(C_{k-1}^{nat}, q_k)$ désigne la réponse du modèle à la question q_k étant donné le contexte C_{k-1}^{nat}.

— **stratégie Oracle** : toutes les réponses sont remplacées par les vraies réponses issues du corpus de questions-réponses considéré r_i^*, indépendamment du comportement du modèle :

$$C_n^{oracle} = \{(q_1, r_1^*), (q_2, r_2^*), \ldots, q_n\}$$

Cette distinction vise à mesurer l'effet de la propagation des erreurs sur le comportement du modèle et des sondes. Deux protocoles sont mis en œuvre pour faire varier la taille du contexte, l'un mesuré en nombre de tours et l'autre en nombre de tokens, afin de capturer deux dimensions complémentaires de l'allongement du contexte.

3.2.2 Génération des contextes

Contexte jusqu'à 400 tours Un premier protocole fait varier le nombre de tours n de 1 à 400, ce qui correspond approximativement à 13 000 ($\pm$1 000) tokens. Un point de mesure est collecté à chaque tour, ce qui permet d'obtenir une granularité fine de l'évolution des métriques. En contrepartie, ce protocole est coûteux en données et en calcul, et ne peut pas passer à l'échelle de contextes de très grande taille. Pour chaque valeur de n et chaque stratégie, on génère le contexte C_n, on collecte les activations internes de la couche l du modèle afin d'alimenter les sondes linéaires, et on calcule $p_n = S(h_{t_1}^{(l)}, \ldots, h_{t_k}^{(l)})$. Les métriques de performance sont ensuite calculées pour chaque tour. Ces étapes sont répétées sur l'ensemble du jeu de test, puis sur plusieurs permutations aléatoires de l'ordre des questions, afin de générer des séquences conversationnelles indépendantes et d'obtenir des estimations plus robustes des métriques.

Contexte jusqu'à 100 000 tokens Un second protocole étend l'analyse aux très longs contextes, jusqu'aux limites des fenêtres de contexte des modèles étudiés. Plutôt que de mesurer chaque tour individuellement, on évalue les métriques à des seuils de taille prédéfinis, ce qui réduit significativement le coût de génération tout en permettant d'observer les tendances globales à grande échelle.

Pour chaque seuil de tokens, les métriques sont évaluées sur un ensemble fixe de 100 questions cibles, mises de côté avant la construction du contexte. Ces questions ne font jamais partie du contexte de remplissage : elles sont posées au modèle une fois le seuil atteint, ce qui permet de mesurer l'impact de la taille du contexte sur une cible constante. L'algorithme 1 détaille cette procédure.

Algorithm 1 Évaluation sur contextes longs

Require: Modèle M, sonde S, questions cibles $\{q_1, \ldots, q_{100}\}$, réponses attendues $\{r_1^*, \ldots, r_{100}^*\}$, seuils $\mathcal{T} = \{0, 5\text{k}, 25\text{k}, 50\text{k}, 75\text{k}, 100\text{k}\}$

for chaque seuil $\tau \in \mathcal{T}$ **do**
  Construire C_τ en ajoutant des paires $\langle q_i, M(C_i)\rangle$ jusqu'à atteindre τ tokens
  for $i = 1$ **à** 100 **do**
    $t_i \leftarrow M(C_\tau \oplus q_i)$ ▷ tokens générés, conditionnés au contexte
    $\left(h_{t_1}^{(l)}, \ldots, h_{t_k}^{(l)}\right) \leftarrow$ extraire_activations(M, t_i) ▷ activations des k tokens de la réponse
    $\mathbf{h}_i \leftarrow \left(h_{t_1}^{(l)}, \ldots, h_{t_k}^{(l)}\right)$
    $y_i \leftarrow \mathbf{1}[a_i \in \text{alias}(r_i^*)]$ ▷ étiquetage
    $p_i \leftarrow S(\mathbf{h}_i)$ ▷ probabilité estimée par la sonde
  end for
  $\text{hallu}_\tau \leftarrow \frac{1}{100}\sum_{i=1}^{100}(1 - y_i)$
  $\text{AUCPR}_\tau \leftarrow \text{AUCPR}((p_1, \ldots, p_{100}), (y_1, \ldots, y_{100}))$
end for

3.3 Modèles

Nous sélectionnons des modèles à poids ouverts couvrant diverses architectures et échelles en privilégiant les tailles modérées. Le choix de modèles à poids ouverts est nécessaire afin de pouvoir accéder aux activations et utiliser les sondes linéaires.

Les modèles retenus proviennent de trois éditeurs distincts (Meta, Google et Alibaba) afin de limiter les biais liés à un pipeline d'entraînement spécifique. Il s'agit par ailleurs de familles de modèles largement adoptées, tant dans la littérature récente (Hsieh *et al.*, 2024; Bai *et al.*, 2025; Ferrando *et al.*, 2024) qu'au sein de la communauté, ce qui facilite la comparaison avec des travaux existants. Enfin, ces modèles présentent des fenêtres de contexte très variées, allant de 8K à 256K tokens, ce qui est particulièrement pertinent pour étudier l'impact de ce facteur sur les performances des sondes.

Les modèles choisis sont les suivants : Llama 3 (1B, 3B) avec une fenêtre de contexte allant jusqu'à 8K tokens (Grattafiori *et al.*, 2024), Gemma 3 (1B, 4B, 12B, 27B) avec des fenêtres de 32K à 128K tokens (Team *et al.*, 2025a), et Qwen 3 (4B) avec une fenêtre de 256K tokens (Yang *et al.*, 2025).

3.4 Les sondes linéaires

Une sonde est une fonction qui, pour k tokens, $S : \mathbb{R}^{k\times d} \to [0, 1]$ prédit p, probabilité que la réponse soit correcte à partir des activations internes $h^{(l)} \in \mathbb{R}^d$ du modèle à la couche l. Pour une réponse r composée de k tokens, la prédiction est :

$$p = S\left(h_{t_1}^{(l)}, \ldots, h_{t_k}^{(l)}\right)$$

On utilise des sondes linéaires, c'est-à-dire une régression logistique ; $S_{linear}(h) = \sigma(w^T h + b)$ où σ est la fonction sigmoïde, $w \in \mathbb{R}^d$ et $b \in \mathbb{R}$.

Les sondes sont entraînées uniquement sur des contextes à un seul tour ($n = 1$) (Azaria & Mitchell, 2023; Marks & Tegmark, 2024) et évaluées sur l'ensemble des tours n, afin de déterminer si un protocole d'entraînement minimal permet de généraliser à des contextes multi-tours. L'étude de sondes entraînées sur des contextes de taille variable est laissée à de futurs travaux.

4 Résultats, Discussion

Dans cette section, nous présentons les résultats obtenus en réponse aux trois hypothèses formulées en introduction. Nous examinons d'abord l'évolution du taux d'hallucination en fonction de la taille du contexte conversationnel (section 4.1), en distinguant les deux protocoles expérimentaux : contextes jusqu'à 400 tours et contextes jusqu'à 100 000 tokens. Nous évaluons ensuite les performances des sondes linéaires de détection des hallucinations sur ces mêmes contextes (section 4.2). Pour chacun de ces deux axes, nous comparons la *stratégie Naturelle* à la *stratégie Oracle*, afin de mesurer l'effet de la correction explicite des erreurs sur le comportement des modèles et des sondes.

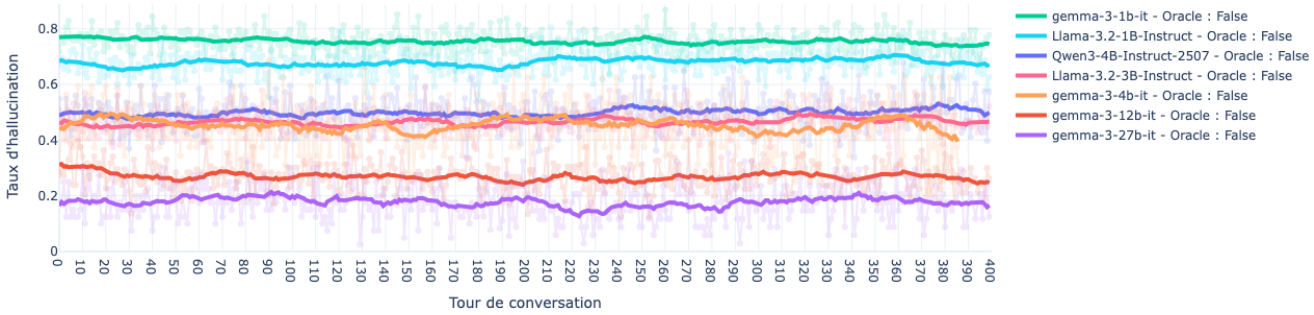

FIGURE 1 – Évolution du taux d'hallucination en fonction du nombre de tours de conversation pour chaque modèle étudié. Les courbes sont lissées par une moyenne mobile de fenêtre 20. Stratégie de génération Naturelle.

4.1 Hallucinations

La figure 1 présente l'évolution du taux d'hallucination par tour pour chaque modèle, sur 400 tours de conversation en stratégie naturelle. On observe que les courbes restent remarquablement stables sur l'ensemble de cet horizon, sans tendance croissante ni décroissante notable, ce qui infirme l'hypothèse H1. De plus, cette stabilité est cohérente entre les trois familles de modèles étudiées. Cependant, les taux d'hallucination varient fortement selon les modèles : gemma-3-27B affiche le taux le plus bas (environ 18%), tandis que gemma-3-1B et Llama-3.2-1B atteignent respectivement 76% et 68%. On note par ailleurs une tendance claire entre la taille du modèle et le taux d'hallucination au sein de chaque famille : plus le modèle est grand, moins il hallucine. C'est ce qu'on attend d'un modèle plus grand et donc capable de stocker plus de connaissances dans ses paramètres. En revanche, la taille de la fenêtre de contexte ne semble pas constituer un facteur déterminant, puisque Qwen3-4B (256K tokens), Gemma-3-4B (128K tokens) et Llama-3.2-3B (8K tokens) présentent tous trois des taux d'hallucination comparables, autour de 45–50%.

La table 1 rapporte les taux d'hallucination moyens pour chaque modèle selon les deux stratégies de génération. On constate que la stratégie Oracle entraîne systématiquement une augmentation du

taux d'hallucination par rapport à la stratégie Naturelle, et ce pour l'ensemble des modèles testés, ce qui infirme l'hypothèse H3. Les écarts varient de 3 points de taux (gemma-3-27B : 0,18 → 0,21) à 11 points (Qwen3-4B : 0,50 → 0,61), sans qu'une tendance claire ne lie l'amplitude de cette dégradation à la taille du modèle. Ce résultat montre que remplacer les réponses du système par les réponses attendues ne bénéficie pas à ses performances ultérieures. Une explication plausible est que les réponses oracle, bien que factuellement correctes, sont hors de la distribution habituelle du modèle (elles ne correspondent ni au style, ni aux connaissances, ni au format que le modèle produirait lui-même), ce qui perturbe ses représentations internes et dégrade sa capacité à répondre correctement aux questions suivantes.

La figure 2 présente l'évolution du taux d'hallucination en fonction de la taille du contexte, jusqu'à 100 000 tokens. Les modèles se montrent globalement robustes à l'allongement du contexte jusqu'à 75 000 tokens, sans dégradation notable du taux d'hallucination, ce qui contraste avec les résultats de (Shi *et al.*, 2023) obtenus sur des modèles plus anciens. Au-delà, le comportement diverge selon les modèles : gemma-3-27B et gemma-3-12B maintiennent des taux stables, tandis que Qwen3-4B et gemma-3-4B présentent une augmentation plus marquée à 100 000 tokens. Cette divergence pourrait s'expliquer par l'influence de la taille d'un modèle dans la gestion des très longs contextes. Concernant l'effet Oracle, on retrouve sur cet horizon la même tendance qu'à 400 tours : la stratégie Oracle dégrade majoritairement les performances (jusqu'à 75K tokens), quel que soit le modèle considéré.

4.2 Sonde de détection des hallucinations

La figure 3 présente l'évolution de l'AUCPR par tour pour chaque modèle sur 400 tours de conversation. Les courbes restent stables sur l'ensemble de cet horizon, sans tendance à la hausse ni à la baisse, ce qui confirme que les sondes entraînées sur tour unique ($n = 1$) généralisent bien aux contextes multi-tours. L'hypothèse H2 est donc infirmée : les sondes ne perdent pas leur capacité de détection à mesure que le contexte croît. Les performances varient fortement selon les modèles, à l'image des taux d'hallucination : gemma-3-27B atteint une AUCPR de 0,92, tandis que les modèles 1B (gemma-3-1B et Llama-3.2-1B) affichent des performances nettement plus faibles, avec des AUCPR de 0,47 et 0,51 respectivement. Ceci suggère que des représentations internes plus riches facilitent la séparation linéaire entre réponses correctes et hallucinées.

La table 1 compare les performances moyennes des sondes selon les deux stratégies de génération.

	Taux d'hallucination		Performances de détection	
Modèle	Stratégie Naturelle	Stratégie Oracle	Stratégie Naturelle	Stratégie Oracle
gemma-3-1B	**0,76**	0,82	**0,47**	0,40
gemma-3-4B	**0,45**	0,55	**0,84**	0,80
gemma-3-12B	**0,27**	0,33	0,85	**0,86**
gemma-3-27B	**0,18**	0,21	**0,92**	0,90
Llama-3.2-1B	**0,68**	0,78	**0,51**	0,41
Llama-3.2-3B	**0,47**	0,53	**0,71**	0,67
Qwen3-4B	**0,50**	0,61	**0,86**	0,84

TABLE 1 – Taux d'hallucination des modèles, suivi de l'AUCPR des sondes de détection d'hallucinations selon la stratégie de génération de contexte. La meilleure valeur est indiquée en gras pour chaque mesure et modèle.

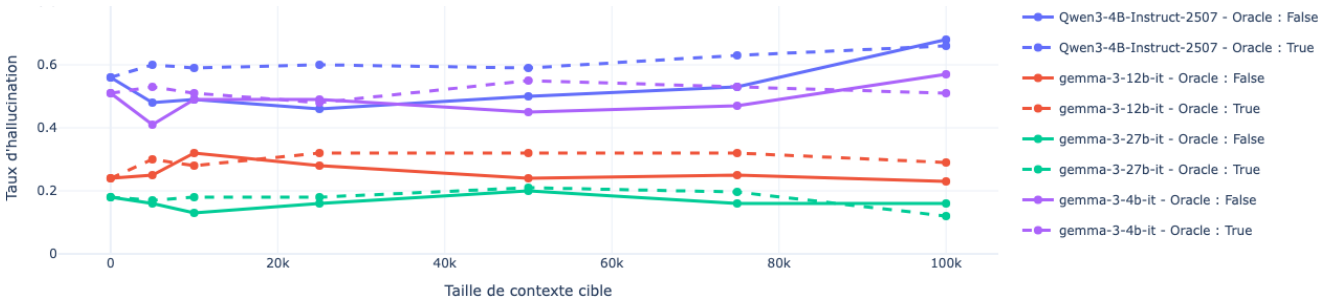

FIGURE 2 – Évolution du taux d'hallucination sur un ensemble de 100 questions cibles, évaluées à des seuils de contexte de 0, 5 000, 25 000, 50 000, 75 000 et 100 000 tokens. Les courbes pleines correspondent à la stratégie Naturelle, les courbes en pointillés à la stratégie Oracle.

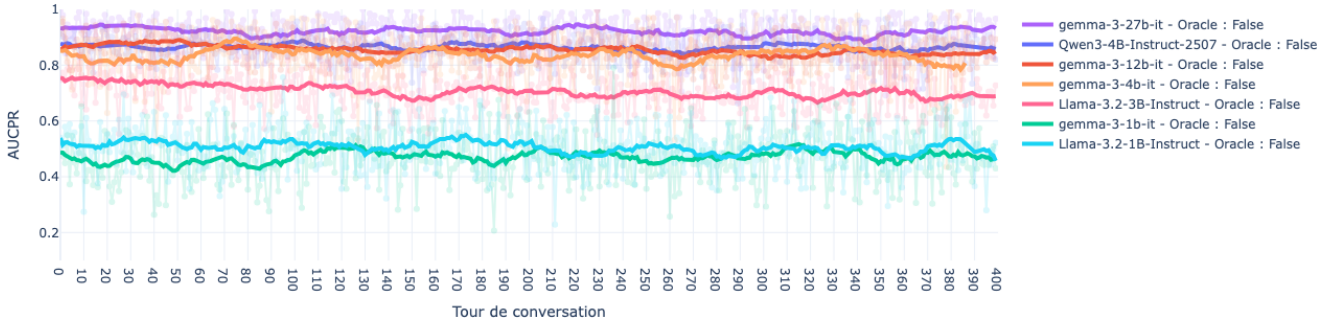

FIGURE 3 – Évolution de l'AUCPR en fonction du nombre de tours de conversation pour chaque modèle étudié. Les courbes sont lissées par une moyenne mobile de fenêtre 20.

La stratégie Oracle dégrade quasi systématiquement l'AUCPR, avec des écarts allant de 2 points (gemma-3-27B : 0,92 → 0,90 ; Qwen3-4B : 0,86 → 0,84) à 10 points (Llama-3.2-1B : 0,51 → 0,41). Seul gemma-3-12B fait exception, avec un léger gain marginal (0,85 → 0,86). Ce résultat est cohérent avec l'effet Oracle observé sur le taux d'hallucination : les réponses oracle, hors distribution du modèle, modifient la structure des activations internes, rendant la séparation linéaire plus difficile pour les sondes.

La figure 4 présente l'évolution de l'AUCPR des sondes en fonction de la taille du contexte, jusqu'à 100 000 tokens. Les résultats sont contrastés selon les modèles. Gemma-3-27B et gemma-3-12B maintiennent des performances stables sur l'ensemble de l'horizon évalué, avec des AUCPR restant au-dessus de 0,85 jusqu'à 100 000 tokens. Gemma-3-4B se distingue des autres modèles par une dégradation plus marquée en stratégie Oracle à partir de 50 000 tokens, tandis que ses performances en stratégie naturelle restent stables sur l'ensemble de l'horizon évalué. Cela suggère que ce modèle est particulièrement sensible à la perturbation induite par les réponses oracle dans les très longs contextes. On retrouve la même tendance que sur 400 tours : la stratégie Oracle dégrade les performances des sondes pour la quasi-totalité des modèles et des seuils évalués.

5 Limites

Plusieurs limites méritent d'être soulignées. Premièrement, le protocole Oracle dans sa forme actuelle (où *toutes* les réponses précédentes sont remplacées par des réponses correctes du corpus) constitue

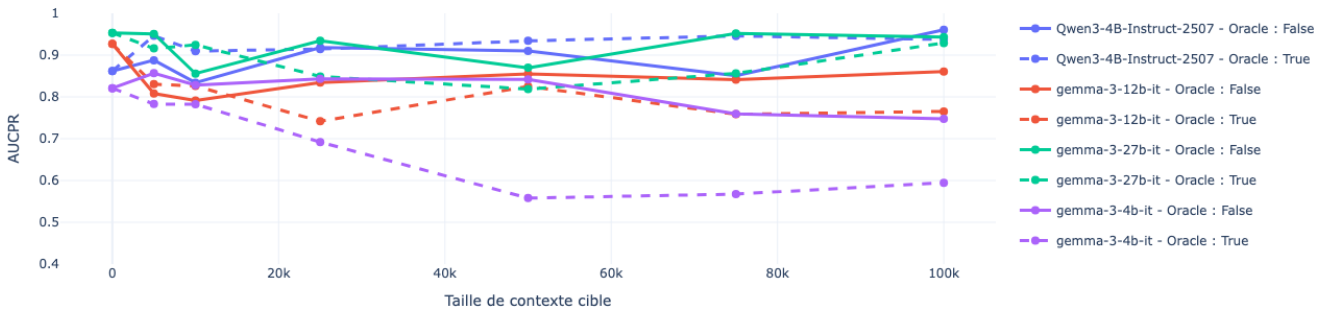

FIGURE 4 – Évolution de l'AUCPR sur un ensemble de 100 questions cibles, évaluées à des seuils de contexte de 0, 5 000, 25 000, 50 000, 75 000 et 100 000 tokens. Les courbes pleines correspondent à la stratégie naturelle, les courbes en pointillés à la stratégie Oracle.

une intervention trop radicale pour isoler proprement l'effet de la correction. Une version plus nuancée, remplaçant sélectivement les réponses fausses ou en éliminant les questions mal répondues du contexte, permettrait de mieux distinguer l'effet de l'accumulation des hallucinations. Deuxièmement, nos expériences n'instruisent pas explicitement les modèles sur la possibilité d'exprimer une incertitude ou de refuser de répondre : bien que les refus soient marginaux dans nos conditions expérimentales, intégrer cette option dans les instructions système pourrait modifier le comportement observé.

6 Conclusion

Ce travail propose une évaluation systématique de l'impact de l'allongement du contexte conversationnel sur le comportement hallucinatoire des LLM et sur les performances des sondes linéaires de détection. Nos expériences, menées sur sept modèles à poids ouverts issus de trois familles distinctes, montrent que les modèles récents sont remarquablement robustes à la saturation du contexte : le taux d'hallucination reste stable jusqu'à 400 tours et jusqu'à 75 000 tokens, infirmant ainsi H1. De même, les sondes entraînées sur tour unique généralisent bien aux contextes multi-tours sans dégradation notable de l'AUCPR, infirmant H2. Ces deux résultats sont encourageants pour le déploiement de systèmes de questions-réponses conversationnels en conditions réelles. En revanche, la stratégie Oracle dégrade systématiquement les performances, tant en termes de taux d'hallucination qu'en termes d'AUCPR, infirmant H3 et suggérant que l'injection de réponses correctes hors distribution perturbe les représentations internes du modèle davantage qu'elle ne les corrige.

Dans un souci de reproductibilité, l'ensemble du code est mis à disposition sur notre dépôt GitLab [1].

Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2025-AD011016140R1 attribuée par GENCI. Ce travail a été financé par l'ANRT dans le cadre du contrat Cifre n°2025/0109 ainsi qu'avec le soutien scientifique et financier du LISN et de SCIAM

1. https://gitlab.lisn.upsaclay.fr/jara/long-context-hallu/-/tree/9c01790b729000186c60597c2583d8bccdc5c505

Références

ALAIN G. & BENGIO Y. (2018). Understanding intermediate layers using linear classifier probes. DOI : 10.48550/arXiv.1610.01644.

ANTHROPIC (2023). Introducing Claude. https ://www.anthropic.com/news/introducing-claude.

AZARIA A. & MITCHELL T. (2023). The Internal State of an LLM Knows When It's Lying. DOI : 10.48550/arXiv.2304.13734.

BAI Y. *et al.* (2025). LongBench v2 : Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. DOI : 10.48550/arXiv.2412.15204.

BELTAGY I., PETERS M. E. & COHAN A. (2020). Longformer : The Long-Document Transformer. DOI : 10.48550/arXiv.2004.05150.

BINKOWSKI J., JANIAK D., SAWCZYN A., GABRYS B. & KAJDANOWICZ T. (2025). Hallucination Detection in LLMs Using Spectral Features of Attention Maps. DOI : 10.48550/arXiv.2502.17598.

CHEN C., LIU K., CHEN Z., GU Y., WU Y., TAO M., FU Z. & YE J. (2023). INSIDE : LLMs' Internal States Retain the Power of Hallucination Detection. In *The Twelfth International Conference on Learning Representations*.

CHOI E., HE H., IYYER M., YATSKAR M., YIH W.-T., CHOI Y., LIANG P. & ZETTLEMOYER L. (2018). QuAC : Question Answering in Context. DOI : 10.48550/arXiv.1808.07036.

DALE D., VOITA E., BARRAULT L. & COSTA-JUSSÀ M. R. (2023). Detecting and Mitigating Hallucinations in Machine Translation : Model Internal Workings Alone Do Well, Sentence Similarity Even Better. 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. 36–50, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.3.

DHULIAWALA S., KOMEILI M., XU J., RAILEANU R., LI X., CELIKYILMAZ A. & WESTON J. (2023). Chain-of-Verification Reduces Hallucination in Large Language Models. DOI : 10.48550/arXiv.2309.11495.

FERRANDO J., OBESO O. B., RAJAMANOHARAN S. & NANDA N. (2024). Do I Know This Entity ? Knowledge Awareness and Hallucinations in Language Models. In *The Thirteenth International Conference on Learning Representations*.

GAO Y., XIONG Y., GAO X., JIA K., PAN J., BI Y., DAI Y., SUN J., WANG M. & WANG H. (2024). Retrieval-Augmented Generation for Large Language Models : A Survey. DOI : 10.48550/arXiv.2312.10997.

GRATTAFIORI A. *et al.* (2024). The Llama 3 Herd of Models. DOI : 10.48550/arXiv.2407.21783.

HSIEH C.-P., SUN S., KRIMAN S., ACHARYA S., REKESH D., JIA F., ZHANG Y. & GINSBURG B. (2024). RULER : What's the Real Context Size of Your Long-Context Language Models ? DOI : 10.48550/arXiv.2404.06654.

HUANG L. *et al.* (2024). A Survey on Hallucination in Large Language Models : Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, p. 3703155. DOI : 10.1145/3703155.

JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y. J., MADOTTO A. & FUNG P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, **55**(12), 248 :1–248 :38. DOI : 10.1145/3571730.

JOSHI M., CHOI E., WELD D. S. & ZETTLEMOYER L. (2017). TriviaQA : A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. DOI : 10.48550/arXiv.1705.03551.

KAMRADT G. (2023). Pressure Testing GPT-4 & Claude 2.1 Long Context. https ://mail.gregkamradt.com/posts/pressure-testing-gpt-4-claude-2-1-long-context.

KOSSEN J., HAN J., RAZZAK M., SCHUT L., MALIK S. & GAL Y. (2024). Semantic Entropy Probes : Robust and Cheap Hallucination Detection in LLMs. DOI : 10.48550/arXiv.2406.15927.

LEWIS P. *et al.* (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. DOI : 10.48550/arXiv.2005.11401.

LIU S. *et al.* (2025). Towards Long Context Hallucination Detection. In *Findings of the Association for Computational Linguistics : NAACL 2025*, p. 7827–7835, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-naacl.436.

MANAKUL P., LIUSIE A. & GALES M. J. F. (2023). SelfCheckGPT : Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. DOI : 10.48550/arXiv.2303.08896.

MARKS S. & TEGMARK M. (2024). The Geometry of Truth : Emergent Linear Structure in Large Language Model Representations of True/False Datasets. DOI : 10.48550/arXiv.2310.06824.

MIN S., KRISHNA K., LYU X., LEWIS M., YIH W.-T., KOH P. W., IYYER M., ZETTLEMOYER L. & HAJISHIRZI H. (2023). FActScore : Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. DOI : 10.48550/arXiv.2305.14251.

OPEN AI (2022). Introducing ChatGPT. https ://openai.com/index/chatgpt/.

REDDY S., CHEN D. & MANNING C. D. (2019). CoQA : A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, **7**, 249–266. DOI : 10.1162/tacl_a_00266.

SHI F., CHEN X., MISRA K., SCALES N., DOHAN D., CHI E., SCHÄRLI N. & ZHOU D. (2023). Large Language Models Can Be Easily Distracted by Irrelevant Context. DOI : 10.48550/arXiv.2302.00093.

TEAM G. *et al.* (2025a). Gemma 3 Technical Report. DOI : 10.48550/arXiv.2503.19786.

TEAM K. *et al.* (2025b). Kimi Linear : An Expressive, Efficient Attention Architecture. DOI : 10.48550/arXiv.2510.26692.

YANG A. *et al.* (2025). Qwen3 Technical Report. DOI : 10.48550/arXiv.2505.09388.