The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Pourquoi les modèles de langage hallucinent, et pourquoi notre manière de les évaluer entretient le problème

Les fausses réponses confiantes que nous appelons "hallucinations" ne sont pas une panne mystérieuse: certaines sont un sous-produit statistique de l'entraînement, et elles persistent parce que les benchmarks dominants récompensent une supposition assurée plus qu'un honnête "je ne sais pas". Une étude de cas sur quatre modèles de pointe montre qu'annoncer les règles de notation dans le prompt (des "rubriques ouvertes") inverse cet incitatif.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · AI-assisted, human-reviewed

Paper: *Evaluating large language models for accuracy incentivizes hallucinations*, Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang, Nature 653, 1047–1050 (2026) · DOI: 10.1038/s41586-026-10549-w

Pourquoi les modèles devinent, et pourquoi nous leur avons appris à le faire

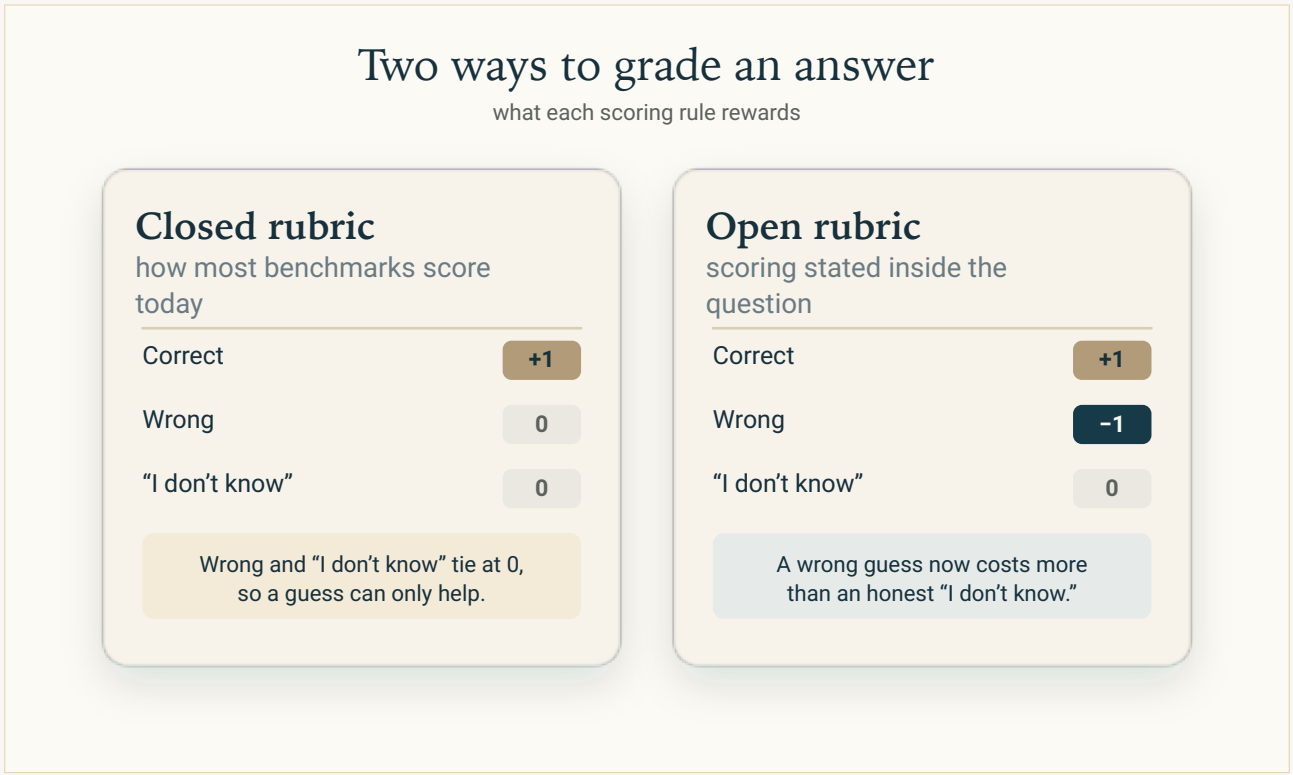
Demandez à un grand modèle de langage la date de naissance d'un inconnu, et il peut répondre "7 mars" avec l'assurance tranquille de quelqu'un qui la lit sur une fiche — tout en se trompant, trois fois de suite, avec trois dates différentes. Les auteurs donnent exactement ce type d'exemple: des modèles de premier plan à qui l'on pose une question factuelle simple — la date de naissance d'une personne, ou le sens d'un acronyme obscur — inventent chacun une réponse différente avec confiance, et aucune n'est correcte. Le mot de l'industrie pour cela est *hallucination*, ce qui donne l'impression d'un défaut de perception. Le premier geste de l'article est d'enlever le mystère.

Commençons par la manière dont un modèle est construit. Dans sa première et plus grande phase d'entraînement, il apprend, en pratique, à quoi ressemble un langage fluide en lisant une quantité énorme de textes. Prenez maintenant un fait sans motif derrière lui — la date de naissance d'une personne précise. Si cette date est apparue une seule fois dans les textes d'entraînement, ou jamais, il n'y a rien à quoi un apprenant de motifs puisse s'accrocher: du point de vue du modèle, la réponse est arbitraire. Les auteurs rendent cela précis en empruntant une vieille idée (celle d'Alan Turing, venue d'un autre problème): si une date de naissance sur cinq n'apparaît qu'une seule fois dans les données, il faut s'attendre à ce qu'un modèle en rate au moins une sur cinq — non parce qu'il est cassé, mais parce qu'il n'y avait jamais rien à apprendre. (Par la même logique, les modèles se trompent presque jamais sur la capitale d'un pays: ces faits apparaissent constamment.) Ils soutiennent, avec prudence, que distinguer une affirmation vraie d'une fausse affirmation plausible est déjà un problème difficile, et que produire *seulement* des affirmations vraies est au moins aussi difficile. Un plancher d'erreur est intégré au système.

> **L'idée dessous: comment compter ce que l'on n'a pas encore vu**
>
> Cela repose sur une idée vraiment élégante — plus ancienne que les modèles de langage, et qui mérite d'être rencontrée correctement.
> **Commencez avec un sac de boules colorées.** Vous ne savez pas combien de couleurs il contient. Vous en tirez 100, une par une, et vous les comptez: rouge 40, bleu 25, vert 15, jaune 5, violet 3, orange 2 — puis **dix couleurs différentes qui n'apparaissent chacune qu'une seule fois.**
> Voici la question que Turing affrontait réellement, dans un tout autre problème: *quelle est la probabilité que la prochaine boule soit d'une couleur que vous n'avez pas encore vue?* Vous ne pouvez pas compter ce que vous n'avez jamais tiré — mais vous **pouvez** compter les couleurs que vous avez vues exactement une fois, les "singletons". L'astuce, appelée **estimation de Good-Turing**, est que la part de vos tirages qui sont des singletons estime la probabilité encore cachée dans les couleurs que vous n'avez pas vues. Dix de vos cent tirages étaient des couleurs vues une seule fois, donc la chance que la prochaine boule soit d'une couleur toute nouvelle est d'environ **10 / 100 = 10%.**
> Ces couleurs vues une fois ne sont pas des *erreurs*. Elles mesurent votre propre ignorance: quand beaucoup de couleurs apparaissent une seule fois, c'est la manière dont l'échantillon vous dit que le monde contient davantage de choses que vous n'avez tout simplement pas

encore tirées.
Remplacez maintenant les couleurs par des dates de naissance, et le sac par le texte d'entraînement du modèle. Supposons que, parmi les dates de naissance qu'il a vues, **une sur cinq apparaisse exactement une fois.** Même astuce: environ un cinquième de la probabilité vit dans des dates que le modèle n'a, en pratique, jamais vues — et une date de naissance n'offre aucun motif de repli (on ne peut pas *déduire* la date de naissance de quelqu'un). Donc une date vue une fois, ou jamais, est une pièce que le modèle ne peut pas pondérer, et il se trompera sur environ **une sur cinq** d'entre elles. Aucune intelligence ne corrige cela: il n'y avait rien à apprendre.
Voilà tout l'argument en miniature: le taux de singletons mesure quelle part du monde est inapprenable *à partir de ces données*, et cela devient un **plancher** sous les erreurs. C'est aussi pourquoi un modèle rate presque jamais une capitale — *Paris* apparaît constamment, son taux de singletons est proche de zéro, donc il y a beaucoup à apprendre.

Cela explique d'où viennent les hallucinations. Cela n'explique pas pourquoi elles *survivent* — pourquoi les modèles, après tout l'entraînement ultérieur censé les rendre utiles et honnêtes, bluffent encore au lieu d'admettre le doute. Ici, l'analogie de l'article est presque inconfortablement juste. Imaginez un élève à un examen qui ne connaît pas la réponse. Si une copie blanche vaut zéro et qu'une tentative peut valoir un point, le mouvement qui maximise la note est de deviner — avec assurance, précisément, jamais "je ne suis pas sûr". Les élèves apprennent cela. Les modèles aussi, apparemment — parce que nous les notons de la même manière. Les auteurs ont passé en revue les benchmarks sur lesquels le domaine se mesure réellement, les classements que les modèles sont réglés pour grimper, et ont trouvé que presque tous donnent exactement la même note à "je ne sais pas" qu'à une mauvaise réponse: zéro. Avec cette règle, un modèle qui devine toujours battra un modèle autrement identique qui signale honnêtement son incertitude. Nous les entraînons, assez littéralement, à le faire par notre système de score.



Les benchmarks peuvent rendre la supposition rationnelle. Avec la notation utilisée par la plupart des benchmarks (à gauche), une mauvaise réponse et un honnête “je ne sais pas” valent tous les deux zéro — donc deviner ne peut qu’aider. Énoncez les règles dans la question, avec une pénalité pour l’erreur (à droite), et s’abstenir quand on n’est pas sûr peut devenir le meilleur choix. Cela change ce que le test récompense; cela ne résout pas, à lui seul, l’hallucination.

C’est la partie à retenir, parce qu’elle va contre le titre habituel. L’hallucination est souvent vendue comme une limite inévitable, presque mystique, de la technologie. L’article conteste les deux idées. Le plancher de préentraînement n’a rien de mystérieux — c’est une erreur statistique ordinaire, du type que l’apprentissage automatique comprend depuis des décennies. Et la persistance n’est pas inévitable — c’est, en partie, un incitatif que nous avons construit et que nous pourrions changer. Un système qui refuserait simplement de répondre quand il n’est pas sûr n’hallucinerait pas du tout; si les modèles déployés ne se comportent pas ainsi, c’est parce que nos tableaux de score punissent le refus.

Ce que les auteurs ont fait

L’article a trois parties. D’abord, un argument mathématique selon lequel une certaine hallucination est statistiquement forcée pendant le préentraînement, en montrant que “générer seulement du texte valide” est au moins aussi difficile qu’un problème binaire de classification “cette affirmation est-elle valide?”. Ensuite, un argument — appuyé par une revue de dix benchmarks influents — selon lequel les mesures courantes de type précision récompensent la supposition plutôt que l’abstention. Enfin, une solution proposée et une étude de cas pour la tester: des évaluations à **rubrique ouverte**, où

la notation est énoncée dans la question elle-même (par exemple, “une bonne réponse vaut 1, une mauvaise vaut -1, donc abstenez-vous si vous êtes sûr à moins de 50%”), afin qu’un modèle puisse savoir quand l’honnêteté est récompensée. Ils l’essaient sur quatre modèles de pointe — Gemini 3 Pro de Google, GPT-5 d’OpenAI, Grok 4 de xAI et Claude Opus 4.5 d’Anthropic — avec les 4 326 questions factuelles de SimpleQA. Ils précisent que l’étude de cas est illustrative, “pas une évaluation contrôlée entre modèles” (réglages par défaut, pas d’ajustement, pas de normalisation du coût).

Ce qu’ils ont trouvé

- **Le préentraînement force une part d’erreur.** Le taux auquel un modèle produit des faussetés confiantes est borné par le bas par (en gros deux fois) le taux d’erreur du meilleur classificateur “cette affirmation est-elle valide?” construit à partir de lui. Pour les faits sans motif apprenable, ce plancher est au moins le *taux de singletons* — la fraction des faits qui apparaissent exactement une fois dans l’entraînement. Une certaine hallucination est inévitable même avec des données parfaitement propres.
- **La notation récompense concrètement la supposition.** Avec une notation ordinaire correct/incorrect, ne jamais s’abstenir est la stratégie optimale, et la revue des auteurs montre que la grande majorité des benchmarks populaires notent “je ne sais pas” comme simplement faux. Un exemple parlant de leur propre côté: sur le test SimpleQA, la précision brute *favorise* légèrement o4-mini d’OpenAI — qui répond presque à tout et se trompe plus des trois quarts du temps — face à GPT-5-mini, qui fait beaucoup moins d’erreurs parce qu’il s’abstient quand il n’est pas sûr. Le modèle le plus téméraire paraît meilleur au classement.
- **Les rubriques ouvertes inversent l’incitatif (dans leur étude de cas).** Ils testent une mitigation simple de l’hallucination (faire répondre le modèle deux fois et s’abstenir si les deux réponses divergent). Avec la précision standard, la mitigation réduit les erreurs *mais réduit aussi la précision* — donc la métrique décourage son adoption. Avec les rubriques ouvertes, la même mitigation devient gagnante pour les quatre modèles sur une plage de pénalités; et GPT-5-mini — que la précision brute avait pénalisé pour s’être abstenu quand il doutait — passe devant o4-mini une fois la notation explicitement énoncée (n = 4 326 questions par modèle).

Ce que cela signifie probablement

Réduire l’hallucination ne consiste surtout pas à inventer davantage de tests spécialisés pour l’hallucination. Il s’agit de changer la façon dont les benchmarks dominants notent l’incertitude, pour qu’admettre “je ne sais pas” ne soit plus puni. Tant que le tableau de score ne change pas, réduire l’hallucination continuera de coûter des points de précision aux modèles, et donc d’être découragé — c’est pourquoi les auteurs décrivent le problème comme “socio-technique”: en partie une meilleure métrique, en partie la nécessité de faire adopter cette métrique par les classements influents.

Ce que cela ne prouve *pas*

- Cela ne montre **pas** que les rubriques ouvertes corrigent l'hallucination dans le monde réel. L'expérience qui les soutient est une petite étude de cas, volontairement **non contrôlée** — quatre modèles avec réglages par défaut, une mitigation choisie, un test de questions factuelles — destinée à démontrer le *renversement d'incitatif*, pas à classer les modèles ni à prouver une efficacité générale.
- Cela ne prétend **pas** que la notation est la *seule* cause. Les erreurs dans les données d'entraînement, les problèmes réellement difficiles et les prompts inconnus restent des sources séparées.
- Cela ne soutient **pas** l'idée populaire selon laquelle les hallucinations sont inévitables. Les auteurs défendent l'inverse: un système qui répondrait seulement aux questions vérifiables et dirait sinon "je ne sais pas" n'hallucinerait jamais.
- Cela ne fait **pas** disparaître le plancher de préentraînement — cela l'explique et le borne, et cette borne concerne les erreurs factuelles confiantes, pas tout le comportement du modèle.
- Cela ne montre **pas** que les rubriques ouvertes suffisent *à elles seules*. Elles changent ce qu'une évaluation récompense; elles ne remplacent pas la recherche d'information, l'usage d'outils ni des modèles mieux calibrés.

Quelle est la solidité de la preuve?

- Le coeur est **mathématique** — des bornes inférieures formelles, pas des mesures. Comme argument théorique, il tient sur ses propres hypothèses.
- Il repose sur des modèles du problème **délibérément simplifiés**; les auteurs signalent eux-mêmes la "fausse trichotomie" qui consiste à traiter toute réponse comme correcte, incorrecte ou "je ne sais pas", ainsi que le cadre idéalisé des "faits arbitraires" utilisé pour la borne la plus nette.
- La revue des benchmarks est un **petit échantillon choisi** — dix évaluations influentes, pas un audit exhaustif.
- L'étude de cas est **réelle mais limitée**: quatre modèles de pointe, une seule mitigation, SimpleQA seulement, réglages par défaut, explicitement "pas une évaluation contrôlée". C'est une preuve de concept pour l'argument d'incitation, pas un résultat de benchmark.
- Il faut nommer le point de vue: trois des quatre auteurs sont ou étaient employés par **OpenAI**, et l'article défend un changement dans la manière dont le domaine évalue les modèles. C'est une position bien argumentée venant d'une partie intéressée, pas une revue extérieure neutre — à peser, non à écarter. (À son crédit, l'article applique la critique à ses propres modèles, o4-mini et GPT-5-mini, aussi volontiers qu'aux autres.)

Pourquoi c'est important

Cela recadre un problème très chargé en hype. L'"hallucination" est souvent vendue soit comme un défaut inquiétant, soit comme un mur infranchissable; cet article la rend ordinaire et en partie auto-

infligée — un plancher statistique que l'on peut réellement comprendre, posé sur un incitatif que nous avons choisi. La leçon plus large est plus discrète et plus utile: les progrès en fiabilité dépendront peut-être autant de *ce que nous mesurons* que de ce que nous construisons.

Résumé propre

Les fausses réponses confiantes des modèles de langage viennent de deux endroits. Le premier est statistique: quand un fait n'offre aucun motif à apprendre, un modèle entraîné à imiter le langage se trompera parfois, et ce plancher peut être estimé (par exemple à partir du nombre de faits qui n'apparaissent qu'une seule fois dans l'entraînement). Le second est incitatif: presque tous les benchmarks qui classent les modèles notent "je ne sais pas" comme une mauvaise réponse, donc deviner gagne toujours — au point qu'un modèle qui se trompe trois quarts du temps peut dépasser un modèle plus honnête qui s'abstient. La proposition des auteurs n'est pas un autre test d'hallucination, mais des "rubriques ouvertes": énoncer la notation dans la question. Dans une étude de cas sur quatre modèles de pointe, cela inverse l'incitatif, de sorte qu'une méthode qui réduit l'hallucination est récompensée au lieu d'être pénalisée. C'est un article théorie-plus-revue avec une petite expérience explicitement non contrôlée, évalué par les pairs dans *Nature*; la correction est prometteuse mais pas encore démontrée à grande échelle, et les hallucinations sont présentées comme ni mystérieuses ni strictement inévitables.

Sans langue de bois

Ce que l'article montre: Une borne inférieure mathématique selon laquelle une certaine hallucination est forcée pendant le préentraînement (au moins le "taux de singletons" pour les faits sans motif); une revue montrant que la plupart des benchmarks importants ne donnent aucun crédit à "je ne sais pas"; et une étude de cas sur quatre modèles dans laquelle annoncer la notation dans le prompt ("rubriques ouvertes") fait gagner une méthode de réduction de l'hallucination, alors que la précision simple la pénalisait.

Ce qui est plausible mais non prouvé: Que les rubriques ouvertes, intégrées aux benchmarks dominants, réduiraient sensiblement l'hallucination dans les modèles déployés. L'expérience de soutien est petite et explicitement non contrôlée.

Ce que cela ne montre pas: Que les hallucinations sont inévitables (l'article soutient l'inverse); que la notation est la seule cause; que l'hallucination peut être éliminée purement et simplement; que l'étude de cas classe les quatre modèles entre eux.

Principales limites: Un modèle délibérément simplifié correct/incorrect/"je ne sais pas" (les auteurs parlent d'une "fausse trichotomie"); une petite revue choisie de benchmarks (dix évaluations); une étude de cas non contrôlée (quatre modèles, une mitigation, un test, réglages par défaut); et un argument porté par OpenAI sur la manière dont le domaine devrait évaluer les modèles.

Quel niveau de confiance pour un lecteur général? Élevé sur le fait que l'hallucination n'est ni mystérieuse ni strictement inévitable, et que les benchmarks dominants récompensent aujourd'hui la supposition. Modéré sur le fait que la solution proposée aide — elle a maintenant une vraie preuve de concept, mais pas encore une démonstration de fonctionnement large et à grande échelle.

Sources

Nature 653, 1047–1050 (2026)
https://doi.org/10.1038/s41586-026-10549-w

Why Language Models Hallucinate (arXiv 2509.04664)
https://arxiv.org/abs/2509.04664

hallucinations-paper-experiments (OpenAI)
https://github.com/openai/hallucinations-paper-experiments

SimpleQA
https://github.com/openai/simple-evals