BASE DE DONNEES PROBANTES - COMPAISS

Taux d'hallucination dans les systemes d'IA contemporains (2023-2025)

Sources verifiees uniquement : etudes evaluees par des pairs, references officielles et rapports de recherche institutionnelle | Prepare aux fins d'evaluation institutionnelle | Toutes les sources verifiees en avril 2026

50-83 %

Taux d'hallucination dans une etude sur des vignettes cliniques portant sur 6 grands modeles (Mount Sinai / Nature Medicine, 2025)

17-34 %

Taux d'hallucination dans des outils juridiques RAG specialises commercialises comme 'sans hallucination' (Stanford HAI, 2024)

33-51 %

Taux d'hallucination des modeles de raisonnement o3 d'OpenAI sur des references factuelles ouvertes (SimpleQA / PersonQA, 2025)

Constat central de l'ensemble des donnees examinees : L'hallucination n'est pas une limite temporaire que l'amelioration des modeles permettra d'eliminer. Il s'agit d'une propriete structurelle persistante des systemes d'IA a inference prioritaire, dont la forme evolue a mesure que les modeles progressent. Sur les taches contraintes et ancrees, les taux diminuent. Sur le raisonnement complexe, les questions en domaine ouvert et les entrees adversariales, les taux demeurent eleves - et dans certains cas augmentent a mesure que les modeles deviennent plus performants.

BASE DE DONNEES PROBANTES - ETUDE PAR ETUDE

Omar et al. (2025) - Attaques adversariales par hallucination en aide a la decision clinique

CLINIQUE / NATURE MEDICINE

MODELES TESTES

GPT-4o, Distilled-DeepSeek-Llama, Phi-4, Gemma-2-27B-it, Qwen-2.5-72B, Llama-3.3-70B (6 modeles, 5 400 sorties)

DEFINITION ET METHODE

300 vignettes validees par des medecins, chacune contenant un detail clinique fabrique (valeur de laboratoire, signe ou affection). Hallucination = le modele elabore sur le detail fabrique comme s'il etait reel.

TAUX RAPPORTES

Parametre par defaut global : 65,9 %
DeepSeek-Llama : 80-83 %
GPT-4o (meilleur cas) : 50-53 %
Avec invite d'attenuation : 44,2 %
GPT-4o + attenuation : 23 %

Evalue par des pairs dans Communications Medicine (Nature Publishing Group), PMCID : PMC12318031 | pmc.ncbi.nlm.nih.gov/articles/PMC12318031 | Source verifiee en avril 2026

Magesh et al. (2024) - Hallucination dans les outils juridiques d'IA RAG specialises

JURIDIQUE / STANFORD HAI

MODELES TESTES

Lexis+ AI (LexisNexis), Westlaw AI-Assisted Research (Thomson Reuters), Ask Practical Law AI (Thomson Reuters), GPT-4

DEFINITION ET METHODE

Plus de 200 requetes juridiques ouvertes construites manuellement. Hallucination = enonce juridique incorrect ou citation mal ancree ne soutenant pas l'affirmation formulee.

TAUX RAPPORTES

Westlaw AI-Assisted Research : >34 %
Lexis+ AI : >17 %
Ask Practical Law AI : >17 %
GPT-4 usage general (etude anterieure) : 58-82 %

Stanford RegLab et HAI | hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-queries | Source verifiee en avril 2026
Constat cle : meme les outils juridiques RAG specialises commercialises comme 'sans hallucination' ont produit des informations incorrectes dans plus d'une requete sur six. La RAG reduit les erreurs mais ne les elimine pas.

Graffius (2026) - Les hallucinations de l'IA s'ameliorent-elles ou s'aggravent-elles?

MULTIMODELE / SYNTHESE DE REFERENCES

MODELES TESTES	DEFINITION ET METHODE	TAUX RAPPORTES
OpenAI o3, o1; Gemini-2.0-Flash; plusieurs fournisseurs. References : Vectara HHEM, SimpleQA, PersonQA, evaluations domaine-specifiques.	Synthese des donnees de references 2024-2025. Deux tendances divergentes : taches de synthese ancrees (taux en baisse) vs raisonnement complexe en domaine ouvert (taux en hausse).	**Synthese ancree, meilleurs modeles : 0,7-1,5 %** **OpenAI o3 sur PersonQA / SimpleQA : 33-51 %** **OpenAI o1 (generation precedente) : ~16 %** **Evaluations de taches larges 2025 : 3-20 %+**

DOI : 10.13140/RG.2.2.33179.53285 | scottgraffius.com/blog/files/ai-hallucinations-2026.html | Source verifiee en avril 2026

Constat cle : les modeles de raisonnement de nouvelle generation hallucinatent davantage sur les taches factuelles complexes que les modeles anterieurs. Une capacite de raisonnement accrue ne se traduit pas automatiquement par une meilleure fiabilite factuelle.

Chelli et al. (2024) - Taux d'hallucination dans les citations academiques generees par l'IA

CLINIQUE / JMIR

MODELES TESTES	DEFINITION ET METHODE	TAUX RAPPORTES
GPT-3.5 (ChatGPT), GPT-4, Bard/Gemini	Reference generee ne correspondant a aucune metadonnee de publication reelle. Revue systematique des citations generees par l'IA pour la litterature medicale.	**GPT-3.5 : 39,6 %** **GPT-4 : 28,6 %** **Bard/Gemini : 91,4 %**

Journal of Medical Internet Research, 2024, e53164 | jmir.org/2024/1/e53164

Tableau de classement Vectara HHEM (2025) - Reference de synthese ancree

REFERENCE

MODELES TESTES	DEFINITION ET METHODE	TAUX RAPPORTES
Gemini-2.0-Flash, GPT-4o, GPT-o1, DeepSeek-R1, DeepSeek-V3 et autres	Hughes Hallucination Evaluation Model (HHEM) : fidelite de la sortie du modele par rapport a un document source fourni. Tache de synthese ancree et contrainte - conditions optimales.	**Gemini-2.0-Flash : ~0,7 %** **GPT-o1 / GPT-4o : ~0,9-1,9 %** **DeepSeek-R1 : 14,3 %** **DeepSeek-V3 : 3,5 %**

Tableau de classement Vectara | github.com/vectara/hallucination-leaderboard

Remarque : il s'agit de taux optimaux sur une tache de synthese contrainte - non representatifs des performances en domaine ouvert ou en raisonnement complexe.

CE QUE LES DONNEES ETABLISSENT

Le dossier de recherche couvrant 2023-2025 etablit trois constats directement pertinents pour la gouvernance institutionnelle de l'IA.

Premierement, les taux d'hallucination dependent de la tache, non du modele.

La meme famille de modeles qui atteint moins de 1 % sur une tache contrainte de synthese documentaire peut produire des erreurs dans 33 a 51 % des requetes factuelles en domaine ouvert. Le tableau de classement Vectara montre que les meilleurs modeles approchent 0,7 % sur la synthese ancree. La fiche systeme d'OpenAI pour o3 documente des taux d'erreur de 33 a 51 % sur SimpleQA et PersonQA. Ces resultats ne sont pas contradictoires - ils refletent la meme architecture sous-jacente fonctionnant dans des conditions differentes. Les institutions qui evaluent l'IA sur la base de references fournies par les fournisseurs observent generalement des performances optimales dans des conditions contraintes, et non les performances sur les types de questions complexes et specifiques aux politiques que leurs utilisateurs poseront reellement.

Deuxiemement, la RAG reduit l'hallucination mais ne l'elimine pas.

L'evaluation independante par Stanford HAI des principales plateformes d'IA juridique - des outils commercialises comme 'sans hallucination' par LexisNexis et Thomson Reuters - a revele des taux d'hallucination de 17 a 34 % dans des conditions reelles de

requetes juridiques. Il s'agit de systemes RAG specialises avec des bases de donnees juridiques organisees. Le constat n'est pas que la RAG echoue - elle reduit substantiellement les erreurs par rapport aux modeles polyvalents - mais qu'elle n'offre aucune garantie architecturale. Les taux d'hallucination dependent de la qualite de la recuperation, de la complexite des requetes et de la presence de fausses premisses, elements qui varient dans les environnements institutionnels reels.

Troisiemement, les modeles plus performants peuvent davantage halluciner sur certaines taches.

La synthese Graffius (2026) documente un phenomene contre-intuitif : la serie o3 d'OpenAI, optimisee pour le raisonnement complexe, hallucine a 33-51 % sur la memorisation factuelle en domaine ouvert - plus du double du taux de la serie o1 anterieure, etabli a environ 16 %. L'etude clinique du Mount Sinai a constate que l'attenuation par invite reduisait le taux d'hallucination adversarial de GPT-4o de 53 % a 23 % - une amelioration significative, mais representant encore un taux d'echec de 23 % sous la meilleure strategie d'attenuation disponible. L'hallucination n'est pas un probleme de qualite de modele que la prochaine generation resoudra. C'est une propriete structurelle des architectures a inference prioritaire.

IMPLICATION POUR LA GOUVERNANCE

Pour les institutions reglementees, la question pertinente n'est pas de savoir si les systemes d'IA actuels hallucinatent a des taux acceptables sur des references controlees. Plusieurs le font. La question est de savoir si les conditions dans lesquelles des taux acceptables sont atteints peuvent etre maintenues de maniere fiable dans un deploiement institutionnel - et quelles sont les consequences lorsqu'elles ne le sont pas.

L'etude juridique de Stanford HAI illustre le mieux cet ecart. Les plateformes qui se reclamaient 'sans hallucination' sur la base de metriques etroites d'exactitude des citations ont produit des informations juridiques incorrectes dans plus d'une requete sur six dans des conditions reelles. Une universite, un hopital ou un organisme gouvernemental qui s'appuie sur les donnees de reference des fournisseurs pour satisfaire ses obligations de gouvernance se fonde sur des chiffres de performance optimaux qui peuvent ne pas tenir dans les conditions reelles d'utilisation institutionnelle.

Une architecture qui conditionne l'existence de l'inference a une autorisation prealable a l'execution ne rivalise pas avec ces systemes sur le taux d'hallucination. Elle supprime les conditions structurelles dans lesquelles l'hallucination se produit pour les requetes non autorisees ou hors portee - non pas en atteignant un taux plus bas, mais en empechant l'inference de s'executer lorsqu'aucune source faisant autorite n'existe pour la soutenir.

REFERENCES VERIFIEES

1. Omar M. et al. (2025). Analyse multi-modele montrant la vulnerabilite elevee des grands modeles de langage aux attaques adversariales par hallucination en aide a la decision clinique. Communications Medicine (Nature). PMCID : PMC12318031. pmc.ncbi.nlm.nih.gov/articles/PMC12318031
2. Magesh V. et al. (2024). Sans hallucination? Evaluation de la fiabilite des principaux outils de recherche juridique par IA. Stanford RegLab / HAI. hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-queries
3. Graffius S.M. (2026). Les hallucinations de l'IA s'ameliorent-elles ou s'aggravent-elles? Nous avons analyse les donnees. DOI : 10.13140/RG.2.2.33179.53285. scottgraffius.com/blog/files/ai-hallucinations-2026.html
4. Chelli M. et al. (2024). Taux d'hallucination et exactitude des references de ChatGPT et Bard pour les revues systematiques. Journal of Medical Internet Research, e53164. jmir.org/2024/1/e53164
5. Vectara (2025). Tableau de classement du Hughes Hallucination Evaluation Model (HHEM). github.com/vectara/hallucination-leaderboard
6. OpenAI (2025). Fiche systeme o3 et o4-mini. Documente des taux d'hallucination de 33-51 % sur les references PersonQA et SimpleQA. Reference et synthetise dans Graffius (2026).