Newsletter N° 16 – Mars 2026

Les hallucinations des IA : pourquoi, comment, et que faire

Le défaut le plus connu des IA génératives reste aussi le plus mal compris.

Vous lui demandez une jurisprudence sur un point précis. Elle vous cite trois arrêts, avec leurs numéros, leurs dates, leurs juridictions. Tout a l'air parfait. Sauf que les arrêts n'existent pas. Pas un seul. Le modèle vient de vous inventer une réalité juridique de toutes pièces, sans le savoir, sans le vouloir, et sans pouvoir s'en empêcher.

Bienvenue dans le monde des hallucinations.

Le phénomène est connu, documenté, moqué. Il a même donné lieu à plusieurs affaires retentissantes, notamment celle de cet avocat américain qui avait remis au juge des conclusions citant des décisions inventées par ChatGPT. Mais malgré la médiatisation, la nature exacte de ces hallucinations reste mal

comprise par la majorité des utilisateurs. Et tant qu'on ne comprend pas comment elles se produisent, on ne peut pas s'en prémunir.

Une hallucination, c'est quoi exactement ?

Le terme est trompeur. Il évoque une défaillance occasionnelle, une sorte de bug. La réalité est plus dérangeante : l'hallucination n'est pas un bug du modèle, c'est une caractéristique structurelle de son fonctionnement.

Pour comprendre, il faut revenir à la nature même d'un grand modèle de langage. Un LLM ne sait pas. Il prédit.

Plus précisément, il calcule, mot après mot, quelle est la suite la plus probable au texte qui le précède. Quand vous lui demandez de citer un arrêt, il ne va pas chercher dans une base de données : il génère une réponse qui ressemble statistiquement à un arrêt. Si dans ses données d'entraînement il a vu beaucoup de citations du type Cour de cassation, chambre sociale, 12 mars 2018, n° 17-12.345, il va produire une citation de cette forme. Mais rien ne garantit que le numéro existe, que la date est juste, ou que l'arrêt a été rendu.

Le modèle ne ment pas, parce qu'il ne connaît pas la vérité. Il produit du texte plausible.

Newsletter N° 16 – Mars 2026

Quand ce texte plausible correspond aux faits, on dit qu'il a raison.

Quand il ne correspond pas, on dit qu'il hallucine. Mais du point de vue du modèle, c'est exactement la même opération.

Plusieurs types d'hallucinations

Toutes les erreurs ne se valent pas. Distinguer les types d'hallucinations aide à mieux les repérer.

Le premier type, le plus visible, ce sont les hallucinations factuelles.

Le modèle invente des faits : une date erronée, une biographie fantaisiste, une statistique inventée, une citation jamais prononcée. Ce sont les plus faciles à détecter, à condition de vérifier.

Le deuxième type, plus pernicieux, ce sont les hallucinations référentielles.

Le modèle invente des sources : un livre qui n'existe pas, un article scientifique fictif, un arrêt inexistant. La forme est impeccable, le contenu est inventé. C'est

dans cette catégorie que sont tombés plusieurs professionnels qui ont remis des travaux contenant de fausses citations.

Le troisième type, le plus subtil, ce sont les hallucinations de raisonnement.

Le modèle suit une logique qui semble correcte, enchaîne des affirmations cohérentes entre elles, mais la conclusion est fausse. Une erreur s'est glissée à un endroit du raisonnement, et tout ce qui en découle est compromis.

Ces hallucinations sont particulièrement dangereuses dans les domaines techniques, juridiques ou médicaux.

Le quatrième type, plus récent, ce sont les hallucinations de contexte.

Le modèle invente ce que vous lui auriez dit auparavant, attribue des propos à votre interlocuteur, ou modifie subtilement le contenu d'un document que vous lui avez fourni. Ce type est particulièrement traître parce qu'il s'appuie sur des éléments réels qu'il déforme.

Newsletter N° 16 – Mars 2026

Pourquoi ne disparaissent-elles pas ?

Avec les progrès rapides des modèles, on aurait pu penser que les hallucinations seraient en voie de disparition. Ce n'est pas le cas. Elles diminuent en fréquence sur certaines tâches, mais elles restent structurellement présentes. Pourquoi ?

Parce qu'elles sont le revers de la médaille de ce qui fait la valeur du modèle. Un modèle qui ne produirait jamais d'hallucination serait un modèle qui refuserait de répondre dès qu'il n'est pas certain. Or, ce que l'on attend d'un LLM, c'est précisément qu'il produise une réponse, même quand l'information

n'est pas dans ses données.

Cette générosité productive est inséparable du risque d'invention.

Les chercheurs travaillent sur des méthodes pour réduire ce risque : entraînement spécifique à dire je ne sais pas, vérification automatique des affirmations, recoupement avec des sources externes. Mais aucune méthode actuelle n'élimine complètement le phénomène. Et il faut s'attendre à ce qu'il reste présent dans toutes les générations à venir, même si son intensité diminue.

Newsletter N° 16 – Mars 2026

Comment les repérer ?

Quelques réflexes pratiques permettent de réduire le risque d'avaler une hallucination.

D'abord, la défiance systématique envers les informations précises. Plus une affirmation est précise (date, chiffre, nom propre, citation, numéro de référence), plus elle mérite vérification.

Le modèle est extrêmement à l'aise pour produire des précisions parce que les précisions abondent dans ses données d'entraînement. Cette aisance n'est pas une garantie de fiabilité.

Ensuite, l'attention au comportement du modèle face à la contradiction. Si vous lui demandez de confirmer une information qu'il a fournie, et qu'il la modifie sans broncher pour vous donner raison, méfiance. S'il vous donne une information puis, quand vous lui demandez la source, en invente une, méfiance accrue. Un modèle qui change facilement de position est un modèle qui n'avait pas de position solide au départ.

Enfin, la vérification croisée. Quand l'enjeu est important, ne jamais se contenter d'une seule source IA. Recouper avec une recherche directe, une base

professionnelle, un confrère. Ce travail de vérification est lourd, mais il est non négociable dans certains contextes.

Comment s'en prémunir efficacement

Au-delà du repérage, plusieurs techniques permettent de réduire le risque d'hallucination en amont.

La première, c'est le prompting structuré.

Demander au modèle de citer ses sources, de signaler les zones d'incertitude, de distinguer ce qu'il sait de ce qu'il suppose. Un prompt explicite du type signale-moi tout point où tu n'es pas certain à plus de 80 pour cent réduit notablement les inventions silencieuses.

La deuxième, c'est le recours au RAG (Retrieval-Augmented Generation).

Cette technique consiste à brancher le modèle sur une base documentaire fiable et à lui imposer de ne répondre qu'à partir de ce qu'il y trouve. Le modèle reste capable d'halluciner, mais le périmètre est réduit. Les outils juridiques sérieux fonctionnent désormais sur ce principe.

Newsletter N° 16 – Mars 2026

La troisième, c'est la décomposition.

Plutôt que de demander une réponse globale, découper la question en sous-tâches vérifiables. Demander au modèle de raisonner étape par étape, puis vérifier chaque étape. C'est plus long, mais beaucoup plus sûr.

La quatrième, c'est la mise en concurrence.

Poser la même question à plusieurs modèles différents (Claude, ChatGPT, Gemini, Mistral) et comparer les réponses. Les hallucinations ont tendance à diverger entre modèles, alors que les informations factuelles convergent.

Le cas particulier du droit

Pour les professionnels du droit, les hallucinations posent un problème spécifique : le formalisme du droit donne aux inventions une apparence de crédibilité que les modèles exploitent particulièrement bien.

Une référence juridique a une forme stéréotypée. Date, juridiction, chambre, numéro de pourvoi, parties. Le modèle a vu des milliers de citations de cette forme dans ses données d'entraînement. Il sait en produire qui sont

indiscernables, à l'oeil nu, de citations authentiques. Seule la vérification dans une base de données peut révéler l'invention.

Concrètement, pour qui exerce dans le droit : ne jamais citer une référence fournie par un LLM sans l'avoir vérifiée dans une source autoritative (Légifrance, Dalloz, LexisNexis, Lamyline). Ne jamais reprendre un raisonnement juridique sans en vérifier les fondements textuels. Considérer l'IA comme un assistant de réflexion, pas comme une source de droit.

Vivre avec, plutôt que contre

Les hallucinations ne sont pas un défaut transitoire des IA actuelles. Elles sont une caractéristique structurelle qui restera présente, à des degrés variables, pendant longtemps. La posture professionnelle adaptée n'est pas d'attendre qu'elles disparaissent, mais d'organiser son travail pour qu'elles soient sans conséquence quand elles se produisent.

Newsletter N° 16 – Mars 2026

Cela veut dire : intégrer systématiquement une étape de vérification, ne jamais utiliser une réponse brute pour une décision engageante, former ses collaborateurs au repérage des inventions, et garder toujours en tête que la fluidité du texte n'est pas un indice de fiabilité, mais simplement un indice de bon fonctionnement du modèle.

Une IA qui hallucine fluidement reste une IA qui fonctionne bien. La fiabilité, c'est nous qui devons l'apporter.

Passez une excellente journée

Gabriel PAPP

gpappAI.com