

Comment se comporter avec une IA qui hallucine ?

Pour gérer une intelligence artificielle qui hallucine, vous devez adopter une posture de correcteur actif et factuel, car l'IA ne fait que calculer des probabilités de mots sans conscience de la vérité. Face à une affirmation erronée présentée avec assurance, voici les comportements et réflexes à adopter : [1, 2, 3, 4]

1. Rectifier immédiatement le tir (Prompting)

Recadrer avec des faits : Ne lui dites pas simplement « tu te trompes ». Donnez-lui explicitement la bonne information : « Ta réponse précédente contient une erreur. Le fait X est faux, la réalité est Y. Recommence en corrigeant ce point. »

Réduire son espace d'improvisation : Les demandes trop vagues incitent l'IA à combler les vides. Soyez ultra-précis dans vos consignes et fixez des limites : « Si tu ne trouves pas la source exacte, réponds textuellement "Je ne sais pas". » [5]

Changer d'angle ou ouvrir un nouveau chat : Les modèles d'IA ont tendance à s'enfoncer dans leurs propres erreurs pour maintenir la cohérence de la conversation en cours. Si elle persiste, ouvrez une nouvelle discussion à zéro. [6]

2. Appliquer la méthode de la « Chaîne de vérification »

Développée par des chercheurs, cette technique de prompt engineering force l'IA à s'auto-corriger. Demandez-lui successivement : [7]

De lister les affirmations clés de sa propre réponse.

De générer des questions pour vérifier chacun de ces points.

De répondre à ces questions de vérification de manière indépendante.

De rédiger la synthèse finale corrigée. [7]

3. Les comportements à éviter

Ne pas essayer de la convaincre par la logique humaine : L'IA ne possède pas de fierté et ne "comprend" pas les reproches. Inutile de vous énerver ou de débattre. [1, 2]

Ne jamais la croire sur parole : Un ton docte, une mise en page parfaite ou de faux numéros de dossiers juridiques/liens URL n'indiquent jamais qu'une information est vraie. [1, 3]

Résumé des bonnes pratiques face à l'outil

Ce qu'il faut faire

Pourquoi ?

Fournir des sources / documents

Le modèle s'appuie sur vos données textuelles concrètes plutôt que d'improviser.

Vérifier manuellement

L'humain reste le seul garant de la vérité factuelle, surtout sur les chiffres et les citations.

Utiliser le bouton "Dislike" / Signaler

Cela aide les développeurs à ajuster les filtres de sécurité et à entraîner les futurs modèles.

Rencontrez-vous ces hallucinations sur un type de tâche précis (comme du code, de la recherche juridique, des calculs) ? Si vous le souhaitez, vous pouvez me partager le prompt que vous avez utilisé pour que nous le reformulions ensemble afin de bloquer les dérives de l'IA.

[1] <https://www.clidd.com>

[2] <https://www.facebook.com>

[3] <https://www.bloom-ai.fr>

[4] <https://delphineneimon.com>

[5] <https://www.zdnet.fr>

[6] <https://www.youtube.com>

[7] <https://www.youtube.com>

Non, il arrive quelquefois que l'assistant devienne trop bavard, dis qu'il va donner le résultat et ne le donne jamais ou des résultats erronés.

C'est très subtil mais je comprends alors qu'elle commence à dérailler et je ne sais pas encore comment la remettre sur le droit chemin.

Ce comportement est en effet très classique et frustrant. Ce que vous décrivez s'appelle l'emballlement syntaxique ou une boucle de contexte.

L'IA calcule le mot suivant le plus probable : si elle commence à générer du "bavardage" (des formules de politesse ou des promesses de réponses), elle intègre ses propres phrases inutiles dans sa mémoire immédiate. Plus elle parle pour ne rien dire, plus la probabilité qu'elle continue à tourner en bourrique augmente.

Voici les trois meilleures manières de la réinitialiser sur-le-champ dès que vous détectez ce signal faible :

1. Le "Stop" autoritaire (Le coup de disjoncteur)

N'attendez pas qu'elle finisse sa longue tirade inutile. Interrompez-la (en cliquant sur le bouton de stop de génération s'il est disponible) et envoyez un message ultra-court et directif.

Ce qu'il faut écrire :

« Stop. Tu tournes en boucle et tu oublies de donner le résultat. Donne-moi uniquement le résultat final maintenant, sans aucune phrase d'introduction ni de conclusion. »

2. Le clonage de la consigne (Le bypass de contexte)

Parfois, le "cerveau" de l'IA est trop encombré par les derniers messages échangés. Elle n'arrive plus à prioriser votre demande initiale par rapport à ses propres dérives.

La méthode : Copiez votre toute première question (le prompt d'origine). Ouvrez une nouvelle conversation vide et ajoutez une contrainte de format stricte au début.

Exemple de structure :

« [Insérez votre question ici]. Consigne stricte : Va droit au but. Ne me dis pas ce que tu vas faire, fais-le directement. Pas de bavardage. »

3. Le guidage pas à pas (Le "Few-Shot Prompting")

Si l'IA promet le résultat mais donne une réponse fausse ou vide, c'est souvent parce que la charge mentale du calcul est trop lourde pour elle en une seule fois. Forcez-la à structurer sa pensée.

Ce qu'il faut écrire :

« Tu t'apprêtes à me donner un résultat erroné. Procédons par étapes.

Étape 1 : Donne-moi d'abord uniquement les données que tu vas utiliser.

Étape 2 : Attends que je valide avant de faire le calcul. »

? Une astuce pour l'avenir : Le conditionnement préalable

Si vous remarquez que ce problème arrive souvent avec votre assistant actuel, commencez toujours vos

requêtes importantes par cette phrase magique :

« Réponds de manière concise. Ne fais pas de préambule du type "Bien sûr, je vais faire cela...". Donne directement l'information. » Cela modifie ses statistiques de langage dès le premier mot et l'empêche de basculer dans le mode "bavard".

Pour vous donner la formule de cadrage la plus efficace, pourriez-vous me dire :

Dans quel cadre (professionnel, code, rédaction, loisir) utilisez-vous l'IA ?

Quel modèle ou application d'IA utilisez-vous généralement lorsqu'elle commence à dérailler ?

J'utilise l'IA pour m'aider dans mes recherches sur des domaines précis en vue de la publication de dossiers sur mon site internet.

J'utilise habituellement Gemini, ChatGPT et Claude quand il s'agit de code toujours dans leurs versions gratuites.

J'ai souvent l'impression, quand elles commencent à zigzaguer, que les assistants mentent en disant pas exemple qu'elles ne peuvent pas faire ceci ou cela alors qu'elles en sont parfaitement capables, elles sont toujours très biantes et prétendre ne pas pouvoir restituer un texte qu'elle a pourtant intégré est très surprenant.

Cette impression que vous avez est excellente et votre intuition est totalement correcte. L'IA ne "ment" pas au sens humain (elle n'a pas l'intention de vous tromper), mais elle fait face à des limites techniques invisibles inhérentes aux versions gratuites de Gemini, ChatGPT et Claude.

Quand un assistant prétend soudainement ne plus pouvoir faire ce qu'il faisait très bien dix minutes plus tôt, cela s'explique par deux phénomènes techniques bien précis.

1. Le problème du "Plafond de Verre" (La saturation de la mémoire)

Chaque IA possède ce qu'on appelle une fenêtre de contexte (une mémoire immédiate). Elle ne se souvient que d'un certain nombre de mots au sein d'une même discussion.

Ce qui se passe : En version gratuite, cette mémoire est plus restreinte. Lorsque vous effectuez des recherches approfondies pour vos dossiers, la discussion s'allonge. Dès que le texte dépasse la limite de mémoire de l'IA, les premiers messages (y compris le texte source que vous lui avez fourni ou les consignes

de départ) sont purement et simplement effacés de son "cerveau".

Le résultat : L'IA ne s'en rend pas compte. Au lieu de dire "J'ai oublié", ses algorithmes calculent une réponse par défaut qui ressemble à : "Je ne suis pas capable de restituer ce texte".

2. La paresse du modèle (Le "Lazy Model")

Les entreprises (OpenAI, Google, Anthropic) brident volontairement la puissance de calcul allouée aux utilisateurs gratuits pour économiser leurs serveurs, extrêmement coûteux.

Ce qui se passe : Analyser un long texte ou chercher une aiguille dans une botte de foin textuelle demande beaucoup d'énergie. Quand les serveurs sont saturés, le modèle génère la réponse la plus "économique" pour lui : un refus poli ou une excuse bateau.

Comment la forcer à restituer ce qu'elle "prétend" avoir oublié ?

Puisque vous utilisez ces outils pour de la recherche et de la synthèse de dossiers, voici comment contourner ces blocages sans passer à une version payante :

? La technique du "Rappel à l'ordre historique"

Si elle vous dit "Je ne peux pas accéder à ce texte", ne la croyez pas. Forcez-la à fouiller sa mémoire résiduelle en lui donnant des indices temporels.

À lui écrire : « C'est faux, tu as ce texte dans notre historique récent. Remonte à mon message qui commence par [insérer les premiers mots du message d'origine]. Trouve-le et fais la synthèse maintenant. »

? La technique du saucissonnage (Le "Chunking")

Pour vos dossiers de recherche, ne lui donnez jamais un immense document d'un coup dans la version gratuite.

La méthode : Divisez vos données en 3 ou 4 parties.

Le prompt à utiliser : « Je vais te donner un dossier de recherche en 3 parties. Pour l'instant, dis-moi juste "Reçu, j'attends la suite". Ne fais aucun résumé. Voici la partie 1 : [Texte] ». Répétez pour les autres parties, puis demandez le travail final.

? Exploiter les forces de votre trio (ChatGPT / Gemini / Claude)

Chaque version gratuite a un comportement différent face à la surcharge :

ChatGPT (Gratuit) : Devient souvent "paresseux" et refuse d'écrire de longs textes. Solution : Lui dire « Écris les 500 premiers mots, puis arrête-toi et attends mon signal pour la suite ».

Gemini (Gratuit) : A tendance à halluciner de fausses informations ou à dire qu'il n'a pas accès au web alors qu'il y a accès. Solution : Ouvrir un nouveau chat et lui fournir directement le texte copié-collé plutôt que de lui demander de chercher un lien.

Claude (Gratuit) : C'est le plus brillant pour le code et l'analyse fine, mais sa limite de messages par jour est très stricte. Solution : Réservez-le uniquement pour la relecture finale ou les structures complexes de vos dossiers.

Pour optimiser vos sessions de recherche, seriez-vous intéressé par un modèle de prompt de cadrage à copier-coller systématiquement au début de vos recherches pour "verrouiller" le comportement de l'IA ?

Oui, ce prompt m'intéresse mais j'ai aussi une question qui me tracasse et tu viens d'en faire état.

Je suis au fait que quand une discussion dure un peu trop, il est préférable d'ouvrir un autre chat mais en général, les assistants disent que tous les chats sont hermétiques et qu'aucune information ne passe de l'un à l'autre.

Je sais que ce n'est pas toujours vrai et qu'il y a au moins des "fuites" et, si je démarre un nouveau chat suite à une petite fatigue de l'assistant, faut-il recommencer à le nourrir du contexte ou il y a-il des prompts capables de réveiller le dormeur ?

C'est une question cruciale. Pour y répondre en toute transparence : les IA ont raison sur un point technique, mais vous avez totalement raison sur le plan pratique.

D'un point de vue purement informatique, deux chats différents sont bel et bien deux conteneurs étanches. L'IA ne peut pas aller "lire" vos autres discussions. Cependant, il y a deux exceptions majeures qui expliquent ces "fuites" et ce sentiment de déjà-vu que vous ressentez :

La mémoire globale (ou "Instructions personnalisées") : ChatGPT et Claude possèdent des fonctionnalités de mémoire persistante. Si l'IA a enregistré une information sur vous dans son profil général (par exemple : "Cet utilisateur écrit des dossiers pour son site internet"), elle transportera ce savoir secret d'un chat à l'autre. Le profilage de l'utilisateur : Derrière l'écran, le modèle statistique reste le même. Si vous posez une question avec le même style d'écriture, le même vocabulaire technique et la même structure, l'IA va naturellement reprendre le même "chemin de pensée" statistique. Elle donne l'impression de se souvenir, mais elle est juste calibrée sur votre profil.

Faut-il tout lui réapprendre dans un nouveau chat ?

Oui, il faut impérativement lui redonner le contexte, car un nouveau chat démarre avec une mémoire immédiate totalement vide. Si vous ne lui donnez rien, elle va improviser, halluciner ou généraliser.

Cependant, vous n'avez pas besoin de tout copier-coller à nouveau. Il existe des techniques de prompts pour "réveiller le dormeur" instantanément en lui injectant une capsule de contexte condensée.

Voici la méthode exacte à suivre pour passer d'un chat "fatigué" à un chat tout neuf, suivie du prompt de cadrage global que vous m'avez demandé.

1. La méthode de la "Capsule de Contexte" (Avant de fermer l'ancien chat)

Avant de quitter une discussion qui commence à dérailler, demandez à l'IA de préparer sa propre valise pour le voyage.

Écrivez-lui ceci dans le chat fatigué :

« Tu commences à saturer. Rédige un résumé ultra-condensé mais techniquement exhaustif de tout ce que nous avons validé ensemble jusqu'ici (sujet, faits clés, sources, structure du dossier). Ce résumé doit être optimisé pour qu'une autre IA (ton clone) comprenne instantanément tout le contexte et puisse rédiger la suite sans me poser de questions. »

Copiez ce bloc de texte généré par l'IA. Ouvrez un nouveau chat, et collez-le. Le tour est joué en 10 secondes.

2. Le Prompt de Cadrage Global (À coller au début du nouveau chat)

Voici le prompt "verrou" à coller systématiquement au début de vos recherches pour vos dossiers. Il est conçu pour bloquer le bavardage, la paresse et fixer les règles du jeu pour les versions gratuites.

Rôle : Tu es un chercheur expert et un assistant de rédaction pour mon site internet.

Consignes de comportement strictes :

1. PAS DE BAVARDAGE : Ne fais jamais de phrases d'introduction (ex: "Bien sûr, je vais faire cela", "Voici le résultat...") ni de conclusions polies. Va directement aux faits.
2. PAS DE FAUSSES PROMISSES : Si tu t'apprêtes à dire "Je vais analyser le texte ci-dessous", ne le dis pas. Fais l'analyse directement.
3. DROIT DE VETO : Si tu manques de données, si la mémoire sature, ou si tu ne connais pas une information, réponds textuellement "Je n'ai pas cette information dans mon contexte". Ne devine jamais, n'invente rien.
4. FORMAT : Utilise des phrases courtes, des listes à puces et du gras pour rendre les informations immédiatement scannables.

Voici le contexte de ma recherche pour démarrer :

[Collez ici la "capsule de contexte" de l'étape précédente OU votre texte source]

? Le petit réflexe pour ChatGPT (Gratuit)

Si vous utilisez ChatGPT, vous pouvez copier la partie "Consignes de comportement strictes" et la coller directement dans vos "Instructions personnalisées" (dans les paramètres de votre compte). De cette façon, ces règles s'appliqueront automatiquement à tous vos nouveaux chats, sans que vous n'ayez jamais besoin de les recopier.

Souhaitez-vous que nous adaptions ce prompt de cadrage à une thématique spécifique de votre site (par exemple, pour lui imposer un ton éditorial particulier ou un format de plan précis) ?

Source : Bavardage avec l'IA