Hallucinations par conception : une évaluation intermodèles de la manière dont les grands cabinets de conseil en IA conseillent leurs clients pour gérer les erreurs de l'IA générative

Aperçu méthodologique

Le présent rapport de synthèse a été élaboré au moyen d'une analyse comparative en plusieurs étapes utilisant quatre grands modèles de langage indépendants - GPT, Perplexity, Copilot et Gemini - auxquels a été fourni un prompt identique et strictement circonscrit, portant exclusivement sur la manière dont les grandes organisations conceptualisent et contrôlent les hallucinations de l'IA. Chaque modèle a produit une évaluation autonome de Deloitte, PwC, KPMG et Microsoft, fondée uniquement sur la documentation publique relative aux produits, aux cadres de gouvernance et aux descriptions architecturales. Ces quatre rapports ont ensuite été traités comme des intrants analytiques indépendants et examinés côte à côte afin d'identifier les zones de concordance complète, d'alignement partiel et de divergence significative. Les conclusions finales présentées ici n'introduisent aucune nouvelle interprétation ni synthèse au-delà de ce qui est étayé par les résultats des quatre modèles; elles reflètent plutôt une évaluation de consensus intermodèles qui pondère les constats en fonction de leur cohérence et de leur récurrence dans l'ensemble des analyses.

Approches de recherche propres à chaque modèle (résumé)

Bien que les quatre modèles aient répondu au même prompt et aux mêmes contraintes de portée, chacun a employé une approche distincte de recherche et d'extraction. Ces différences sont résumées ci-dessous afin d'assurer la transparence quant à la manière dont les évaluations sous-jacentes ont été produites.

GPT

GPT a mené une analyse structurée fondée sur des documents, ancrée dans des documents publics des firmes, y compris des cadres de gouvernance de l'IA, de la documentation sur l'IA responsable, des descriptions de produits et des explications architecturales. L'analyse visait à extraire les hypothèses explicites et implicites relatives aux hallucinations, à les cartographier aux points de contrôle du cycle de vie (avant, pendant ou après la génération) et à classifier le paradigme architectural implicite. GPT n'a pas eu recours à la recherche Web en temps réel lors de la phase finale de synthèse; il a plutôt traité les documents cités dans les rapports fournis comme faisant autorité et a strictement limité ses conclusions à ce que ces documents permettaient d'étayer.

Copilot

Copilot a appliqué une méthodologie de synthèse interrapports plutôt que d'introduire une expansion indépendante des sources. Il a traité les évaluations des firmes générées par les quatre modèles comme des intrants analytiques fixes et a procédé à une normalisation à l'aide

d’une taxonomie commune (cadre de définition des hallucinations, inévitabilité ou évitabilité, point de contrôle du cycle de vie et paradigme architectural). Copilot a mis l’accent sur une pondération du consensus fondée sur les preuves, en distinguant explicitement les niveaux d’accord fort, modéré et mixte, et a évité de résoudre les désaccords par l’interprétation. Aucun nouveau fait ni aucune inférence architecturale n’ont été introduits au-delà de ce qui figurait déjà dans les rapports eux-mêmes.

Perplexity

Perplexity a combiné une recherche Web ciblée et une extraction fondée sur des documents. Il a lancé des requêtes ciblées sur les sites Web des firmes, des livres blancs, des blogues techniques et, dans le cas de Microsoft, de la documentation de plateforme, portant spécifiquement sur les hallucinations, l’ancrage, le RAG et les contrôles d’IA responsable. Il a ensuite recoupé ces constats avec le rapport compilé fourni afin d’en assurer l’alignement. Lors de la phase de consensus, Perplexity a traité le document compilé comme une source de vérité et a restreint la synthèse aux points explicitement présents dans les analyses des quatre modèles, en signalant les niveaux de consensus selon la récurrence plutôt que l’emphase.

Gemini

Gemini a adopté une approche médico-légale axée sur la mise à l’épreuve d’hypothèses, visant à identifier les hypothèses architecturales sous-jacentes plutôt que les affirmations de gouvernance de surface. Il a analysé les cadres des firmes, la documentation des plateformes et, pour Microsoft, les descriptions des outils techniques afin de déterminer le moment où l’intervention se produit par rapport à l’exécution du modèle. Gemini a explicitement vérifié si une organisation démontrait un refus d’exécution avant génération ou un blocage strict de l’inférence et a signalé des constats négatifs lorsqu’aucune preuve de ce type n’apparaissait. Lors de l’étape intermodèles, Gemini a mis l’accent sur l’alignement des interprétations des modèles afin de garantir que les conclusions reflètent une compréhension collective plutôt qu’une lecture propre à un seul modèle.

Synthèses finales de la recherche de chaque modèle d’IA

ChatGPT - Synthèse de l’évaluation des quatre firmes

Position générale

GPT conclut de façon constante que les quatre firmes fonctionnent selon un paradigme de génération d’abord, avec atténuation a posteriori. Les hallucinations sont traitées comme des propriétés inhérentes et probabilistes des LLM et sont gérées au moyen de la gouvernance, du RAG et de la supervision humaine après l’inférence.

Par firme

- Deloitte : Les hallucinations sont présentées comme des défaillances de fiabilité et des risques liés aux modèles; elles sont inévitables et régies après la génération par l'assurance qualité, la gouvernance et la révision.

- PwC : Les hallucinations sont décrites comme des sorties plausibles mais non fondées; elles sont atténuées par le RAG et la discipline des prompts pendant et après la génération, sans être structurellement empêchées.

- KPMG : Les hallucinations sont traitées comme des risques endémiques de fiabilité; elles sont abordées principalement par la gouvernance, l'auditabilité et la surveillance après la génération.

- Microsoft : Les hallucinations sont présentées comme du « contenu non ancré »; elles sont atténuées par l'ancrage, des boucles de correction et la gouvernance après l'inférence.

Constat clé

GPT ne relève aucune preuve de refus d'exécution avant l'inférence ni de prévention structurelle chez aucune des firmes.

GEMINI - Synthèse de l'évaluation des quatre firmes

Position générale

Gemini s'aligne étroitement sur GPT, en mettant l'accent sur les hallucinations comme des artefacts inévitables de modèles probabilistes, gérés par des contrôles en couches après la génération.

Par firme

- Deloitte : Les hallucinations correspondent à un risque de fiabilité des sorties; elles sont inévitables et contrôlées après la génération par la gouvernance et la révision humaine.

- PwC : Les hallucinations sont présentées comme un décalage entre la probabilité et la réalité; elles sont traitées au moyen de « membranes », du RAG et de la révision, mais la génération a toujours lieu.

- KPMG : Les hallucinations sont traitées comme un risque persistant; elles sont régies au moyen de cadres de gestion des risques traditionnels après la génération.

- Microsoft : Les hallucinations sont formalisées comme un manque d'ancrage; elles sont réduites par l'ancrage et la correction, sans être éliminées.

Constat clé

Gemini affirme explicitement qu'aucune des quatre firmes ne remet en question l'hypothèse de la génération d'abord.

COPILOT - Synthèse de l'évaluation des quatre firmes

Position générale

Copilot caractérise les quatre organisations comme supposant que les hallucinations sont attendues et inévitables, et que leur atténuation est principalement axée sur les étapes pendant et surtout après la génération.

Par firme

Deloitte : Les hallucinations constituent des risques opérationnels et de gouvernance; elles sont gérées après la génération par la validation et la supervision.

PwC : Les hallucinations représentent des risques inhérents à l'IA générative; elles sont atténuées par les prompts, le RAG et la gouvernance de l'IA responsable.

KPMG : Les hallucinations sont intégrées aux préoccupations de fiabilité et de confiance; elles sont traitées par des contrôles et des audits post-génération.

Microsoft : Les hallucinations sont considérées comme des défaillances d'ancrage; elles sont atténuées par des outils de plateforme (RAG, correction, couches de sécurité).

Constat clé

Copilot ne relève aucune défense du blocage avant la génération; le refus est traité comme un comportement configurable, et non comme une architecture par défaut.

PERPLEXITY - Synthèse de l'évaluation des quatre firmes

Position générale

Perplexity fournit le niveau de détail le plus granulaire sur le cycle de vie, mais parvient à la même conclusion : les hallucinations sont structurellement inévitables et sont atténuées par des contrôles compensatoires superposés à des systèmes axés sur la génération d'abord.

Par firme

Deloitte : Les hallucinations sont des défaillances de fiabilité statistiquement attendues; elles sont atténuées par le RAG, l'assurance qualité et la gouvernance après la génération.

PwC : Les hallucinations sont des risques inhérents à l'IA générative; elles sont réduites par des techniques de prompting, des « doubles vérifications » par RAG et la révision.

KPMG : Les hallucinations sont présentées comme des risques endémiques de fiabilité; elles sont gérées par la gouvernance, la validation et la surveillance après la génération.

Microsoft : Les hallucinations sont définies comme des sorties non ancrées; elles sont abordées par l'ancrage, la détection et la correction après l'inférence initiale.

Constat clé

Perplexity souligne explicitement que le RAG est traité comme un mécanisme compensatoire, et non comme une barrière stricte empêchant l'inférence.

Observation métamodèle intermodèles (dérivée des quatre synthèses)

Les quatre modèles d'IA s'entendent indépendamment sur les points suivants :

- Les hallucinations sont traitées comme inévitables par les quatre firmes.
- L'inférence se produit par défaut.
- L'atténuation intervient pendant et surtout après la génération.
- Aucune firme n'est décrite comme imposant un refus d'exécution structurel avant l'inférence.