Hallucide

Vérifier l'IA contre les sources officielles.

Hackathon de l'Assemblée nationale 2026, défi « IA et Hallucination ».
Porteur : Hamza Konte. Démonstration : http://141.11.165.40:8770

Le problème

Les IA génératives produisent des réponses inexactes ou non sourcées, et hallucinent jusqu'à leurs propres justifications : le modèle invente une date, un vote, un mandat, puis fabrique une explication convaincante pour le défendre. Plus la réponse est fluide, plus elle inspire confiance, et c'est ce qui la rend dangereuse.

En contexte institutionnel, distinguer une information vérifiée d'un contenu généré est impossible à l'œil nu. Une erreur reprise dans une note, une question écrite ou un article engage la responsabilité de son auteur. L'illusion de fiabilité est le cœur du danger, pas l'erreur elle-même.

La solution

Hallucide est une couche de confiance entre le modèle et l'utilisateur. La réponse est interceptée avant affichage, découpée en affirmations élémentaires, puis chaque affirmation est confrontée à l'open data officiel :

- articles de code consolidés (Légifrance) ;
- questions parlementaires (écrites, orales, au Gouvernement) ;
- mandats et commissions des députés ;
- données tabulaires data.gouv.fr, tracées à la cellule exacte.

La réponse n'est affichée qu'annotée, affirmation par affirmation, avec sa source datée et un lien vers le document officiel. Ce que le système ne peut pas sourcer, il le signale au lieu de le présenter comme un fait.

Le principe : le verdict vient de la source, jamais du modèle

Le moteur de vérification est un pipeline déterministe :

1. La question est décomposée en intentions atomiques.

2. Chaque intention déclenche une récupération directe dans la source officielle (serveur MCP unifié Parlement / Législation, data.gouv.fr).
3. Chaque affirmation est vérifiée mot pour mot contre le passage récupéré. Un texte abrogé n'est jamais présenté comme en vigueur.
4. Un plancher de risque incontournable s'applique : référence inférée, troncature, pertinence non garantie ou couverture insuffisante font monter le risque, sans possibilité de contournement.
5. Tout résultat à risque élevé exige une décision humaine explicite, capturée et journalisée. Rien n'est publié automatiquement dans ce cas.
6. Chaque étape alimente un journal de conformité rejouable, qui ne contient ni la question posée ni l'identité de l'utilisateur.

Le modèle de langage (Claude, Mistral ou Gemini, interchangeables) sert à décomposer et à mettre en forme. Il ne juge jamais sa propre fidélité.

Les statuts affichés

Statut	Sens
Vérifié	Présent mot pour mot dans une source officielle opposable en vigueur
Donnée tracée	Valeur retrouvée à la cellule exacte d'une donnée publique
Prudence	Exact mais non opposable, ou reformulation à valider
Risque	Non retrouvé dans la source : potentielle hallucination, bloqué

Un indice de confiance global accompagne chaque réponse. Il traduit les statuts établis par le code ; ce n'est jamais une probabilité du modèle.

Ce que la démonstration montre

Posez « À quelles commissions fait partie Yaël Braun-Pivet depuis sa dernière législature ? » : chaque appartenance est affichée avec ses dates, son rôle et un lien vers la fiche officielle du député. Une question dont la prémisse est fausse ne produit aucune invention : le système répond qu'il ne peut pas sourcer, plutôt que d'halluciner.

220 tests automatisés couvrent le moteur. Deux scénarios ont été joués en direct contre les sources réelles : prémisse fausse sans hallucination, et passage authentique mais hors sujet bloqué par le plancher de risque.