

Hallucide, design technique

Hackathon « IA et Hallucination », défi Assemblée nationale. Porteur : Hamza Konte.

Nom

Hallucide : *hallucination* + *-cide* (tuer, comme pesticide) + écho *lucide*. Le tueur d'hallucinations qui rend l'IA lucide.

Problème

Les IA génératives produisent des réponses inexactes ou non sourcées, et hallucinent jusqu'à leurs propres justifications : une illusion de fiabilité sans ancrage réel. En contexte institutionnel, distinguer une information vérifiée d'un contenu généré est difficile.

Positionnement

Hallucide est une **couche de confiance par détection et annotation**, fidèle à la consigne : la réponse n'est affichée qu'annotée. Il ne corrige pas le texte du modèle. Ce qu'il tue, c'est **l'illusion de fiabilité** : toute affirmation non ancrée dans une source officielle est visiblement marquée, donc ne peut plus se faire passer pour un fait. L'utilisateur ne reçoit jamais une hallucination estampillée « vraie ».

Principe dur : le verdict vient de la source, jamais du modèle. Le LLM décompose et met en forme ; il ne juge jamais sa propre fidélité.

Choix techniques

- **Modèle : Claude (API Anthropic)** pour la décomposition des questions et la mise en forme des réponses. Providers Mistral API et Gemini API branchés en alternative : l'interface `ModelProvider` les rend interchangeables, aucun ne participe au verdict.
- **Couche source : serveur MCP tricoteuses** interrogé directement (<https://mcp.code4code.eu/mcp>, Streamable HTTP). Aucune ingestion locale, aucune base à maintenir : le backend est client MCP et interroge la donnée officielle à la volée (`query_sql` paramétré sur les schémas

assemblee/legifrance, recherche plein texte, recettes métier).

data.gouv.fr en complément pour les données tabulaires.

- **Backend** : Python stdlib ([http.server](http://server)), zéro dépendance de service.
Deux endpoints : POST /resolve (détection de route) et POST /ask (pipeline complet).
- **Frontend** : chat statique (HTML/CSS/JS sans framework), DSFR pour les composants. Prose annotée, indice de confiance, une ligne par affirmation avec lien vers le document officiel.
- **Déploiement** : VPS unique, service systemd, <http://141.11.165.40:8770>.

Architecture / flux (état réel du code)

Question utilisateur

|

▼

[1] /resolve – détection de route déterministe (regex, pas de LLM)
| code_article / parlement_question / commissions / donnee /
texte_libre

▼

[2] Décomposition – Claude découpe la question en intentions
atomiques
| (routes structurées comme commissions : chemin 100 %
déterministe, sans LLM)

▼

[3] Récupération – l'orchestrateur appelle lui-même le MCP
tricoteuses
| (le LLM ne choisit jamais la source) ; data.gouv.fr pour le
tabulaire

▼

[4] Vérification verbatim – chaque affirmation est comparée mot pour
mot
| au passage officiel récupéré ; contrôle du cycle de vie (texte
abrogé
| jamais présenté en vigueur) ; ancres dures sur négations et
chiffres

▼

[5] Triage – plancher de risque incontournable : référence inférée,
| troncature, pertinence non garantie, couverture insuffisante,
| sélection ambiguë → risque élevé, non contournable par
l'appelant

▼

[6] Validation humaine – tout résultat à risque élevé est marqué
| « non publiable en l'état », décision humaine journalisée

▼

[7] Audit – journal de conformité rejouable, sans la question ni
l'identité

▼

Affichage annoté (jamais de réponse brute non vérifiée)

Statuts établis par le moteur : AUTHENTIFIÉ (verbatim dans une source opposable en vigueur), DONNÉE_TRACÉE (valeur à la cellule exacte), CITÉ_NON_OPPOSABLE, INTERPRÉTATION, NON_AUTHENTIFIÉ, NO_ANSWER (le système se tait plutôt que d'inventer).

Couche de présentation (scores et couleurs)

Le score 0-100 affiché est un **habillage déterministe** du statut du moteur, pas une nouvelle logique de confiance :

Statut moteur	Score	Bande
AUTHENTIFIÉ	96	vérifié (vert)
DONNÉE_TRACÉE	88 à 100, déterministe par affirmation	donnée tracée (bleu)
CITÉ_NON_OPPOSABLE	55	prudence (orange)
INTERPRÉTATION	50	prudence (orange)
NON_AUTHENTIFIÉ	20	risque (rouge)
NO_ANSWER	5	risque (rouge)
Résultat non publié	plafonné à 45	revue humaine requise

L'indice global d'une réponse est la moyenne des scores de ses affirmations ; les arcs du donut restent proportionnels à la longueur de chaque affirmation. Le score d'une intention est celui de son affirmation la plus faible : le maillon faible gouverne.

Interface

- Réponse rédigée par Claude **uniquement à partir des lignes vérifiées** (consigne stricte : aucun fait ajouté), masquée par défaut.
- Une ligne par affirmation datée, triée du plus récent au plus ancien, avec badge de statut, score et lien externe vers le document officiel (fiche député de l'Assemblée nationale, article Légifrance).
- « Voir plus » par paquets de cinq au-delà de dix lignes.
- Journal de conformité replié, consultable par affirmation.
- Sans clé API, l'interface affiche « moteur non connecté » : rien n'est jamais simulé.

Vérification temporelle

- Identité par identifiants stables (PAxxxx pour un député, uid de scrutin), jamais par le titre seul.

- Fait-événement (« a voté pour ») vérifié contre l'enregistrement daté ; fait-état (« siège à la commission ») contre l'état valide à la date.
- Si aucun enregistrement unique et daté ne peut être épinglé, le système affiche « ambigu » avec les candidats datés : il refuse de certifier sa propre certitude.

Scénario démo

« À quelles commissions fait partie Yaël Braun-Pivet depuis sa dernière législature ? » : chaque appartenance sort avec ses dates, son rôle et le lien vers la fiche officielle. Une question à prémisse fausse ne produit aucune invention : le système récupère le vrai texte et répond NO_ANSWER plutôt que d'halluciner.

Risques connus et mitigations

- **Dépendance réseau au MCP tricoteuses** pendant la démo : repli gracieux (« source indisponible » plutôt que crash).
- **Latence API** : le goulot est l'appel LLM ; les routes structurées (commissions, article de code) passent par un chemin déterministe rapide.
- **Licence DSFR** : usage réservé à l'État ; acceptable pour un hackathon organisé par l'Assemblée nationale, en-tête Marianne non utilisé.