

Les biais et les hallucinations

Fiabilité ■ Éthique ■ Audit

💡 1. L'IDÉE ESSENTIELLE

Les modèles de langage ne fonctionnent pas comme des bases de données exactes, mais comme des **moteurs de prédiction probabiliste**. Lorsqu'ils manquent d'informations ou qu'ils reproduisent les défauts de leurs données d'entraînement, ils génèrent des erreurs : des **hallucinations** (fausses informations affirmées avec aplomb) et des **biais** (stéréotypes culturels, sociaux ou logiques).

⚠️ 2. LE MÉCANISME DES HALLUCINATIONS

Une hallucination survient quand l'IA prédit la suite de mots statistiquement la plus plausible... mais factuellement fausse.

- **Absence de vérité terrain** : L'IA cherche à faire des phrases cohérentes, pas à vérifier la réalité physique ou historique.
- **Sur-confiance textuelle** : Le modèle utilise le même ton assuré qu'il s'agisse d'une vérité absolue ou d'une pure invention.
- **Lacunes documentaires** : Face à un sujet rare ou inconnu, le modèle « comble les trous » par association d'idées.

🔍 3. D'OÙ VIENNENT LES BIAIS ?

Les biais proviennent directement de la matière première avec laquelle les modèles sont conçus :

- **Biais de corpus** : Internet reflète les inégalités, stéréotypes et visions culturelles dominantes (notamment occidentales et anglophones).
- **Biais de renforcement (RLHF)** : Les préférences humaines lors de l'entraînement correctif peuvent introduire des biais de conformisme ou lisser excessivement les points de vue.
- **Biais de confirmation** : L'IA a tendance à abonder dans le sens des hypothèses formulées par l'utilisateur dans son prompt.

🔍 4. COMMENT AUDITER ET LIMITER LES ERREURS

Vérification croisée	Exiger des sources ou confronter les dires du modèle avec des bases documentaires fiables (couplage RAG).
Test contradictoire	Poser la question sous un angle opposé pour tester la robustesse et détecter un éventuel biais de complaisance.
Température et réglages	Baisser la « température » du modèle dans les API pour privilégier les choix les plus factuels et déterministes.

💡 À RETENIR

« Une IA ne "sait" pas : elle prédit. Connaître ses biais et ses propensions à halluciner est indispensable pour l'utiliser avec recul et esprit critique. »