Vers des IA génératives *intègres*

Sarah Cohen-Boulakia
Directrice Adjointe, Institut DATAIA
Université Paris-Saclay

Tentative de définition d'une IA générative intègre

Quelques premiers critères

IA générative Fiable

- Résultats dont on *contrôle les* ***biais***
- Verification que les résultats sont corrects

IA générative Transparente

- **Traçabilité** des principales sources utilisées
- Vers une **explicabilité** du contenu généré
- Ouverture des modèles
- **Reproductibilité**

IA générative Responsable

- Des usages contrôlés
  - Cadre **juridique**
  - Cadre **écologique**

IA générative fiable

Les biais des IA génératives

Un premier exemple

Mon fils Michael a 17 ans, Michael aime les sciences et Michael aime expliquer, s'il y avait un seul métier parfait pour lui quel serait-il ?

C'est une belle combinaison : aimer les sciences *et* aimer expliquer ! Si on devait imaginer **un seul métier "parfait"** pour Michael avec ces deux ingrédients, je dirais :

👉 Chercheur-enseignant en sciences

(le combo maître de conférence / professeur d'université)

Pourquoi ?
Parce que ce métier lui permettrait :

- **d'explorer les sciences en profondeur** (recherche)
- **de transmettre sa passion aux autres** (enseignement)
- **de continuer à apprendre toute sa vie**
- **de contribuer à la société par la connaissance**
- et aussi de travailler en équipe, de voyager, de faire des conférences, d'inspirer les jeunes...

Ma fille Lucie a 17 ans, elle aime les sciences et elle aime expliquer, s'il y avait un seul métier parfait pour elle quel serait-il ?

C'est super que Lucie aime à la fois les sciences **et** expliquer — c'est une combinaison puissante ! Si on devait imaginer **un seul métier parfait** pour elle, basé uniquement sur ces deux dimensions, ce serait peut-être :

👉 Chercheuse et médiatrice scientifique

Un métier qui lui permettrait :

- **d'explorer la science à fond** (curiosité, rigueur, recherche, labo, terrain...),
- **et de la transmettre avec passion** (pédagogie, vulgarisation, conférences, formats créatifs...).

chatGPT 4.o

Les biais des IA génératives
un second exemple

Wafaâ AMAL • 2e
CEO Prisme.ai | Activate the power of Gen AI
4 j •

Yesterday I tried the GPT trend of the day...
You tell me — should I laugh or should I cry??!

Before you even wonder, let me stop you right there :

✅ Yes, I'm a woman.

✅ AND (not but) I'm also an IT engineer.

✅ AND (not but) I'm leading a GenAI startup.

So YES, I know how to prompt.

Starter pack

Les biais des IA génératives

- Les IA génératives apprennent (s'entraînent) sur des ensembles de données : textes, images...
- Les IA génératives prédisent le **mot suivant ou le pixel voisin le plus *vraisemblable***
- La génération est fondée sur des ***probabilités***
- Les réponses reflètent les caractéristiques des données d'entrée avec un "effet loupe"

Vers des IA fiables - Débiaiser les IA génératives

Libération

Polémique

IA : Gemini génère des nazis noirs, X s'enflamme et Google suspend la création d'images de personnes en invoquant des «problèmes»

Ce jeudi 22 février, Google a décidé de suspendre la création d'images de personnes sur Gemini. L'IA s'est attiré les foudres d'internautes sur X après avoir généré des images de nazis noirs ou de Vikings asiatiques.

Injecter une correction des biais après la phase d'apprentissage ne donne pas les résultats escomptés …

La gestion des biais est un problème difficile mais de mieux en mieux cerné

Enjeux

- Identifier et mesurer les biais des ensembles de données
- Identifier et mesurer les biais dans les modèles

De beaux **problèmes de recherche interdisciplinaires** à l'interface des mathématiques et de l'informatique

IA générative transparente

Les IA génératives ne sont pas explicables

- Elles reposent sur des techniques de Deep Learning (réseaux de neurones profonds)
- Elles agissent en mode "boite noire"
- Leurs décisions (générations) sont fondées sur un modèle probabiliste

Et pourtant...

- On entend parler d'IA transparentes
- et d'IA qui citent leurs sources...


Columbia Journalism Review.

CJR PARTNER TOW CENTER

AI Search Has A Citation Problem

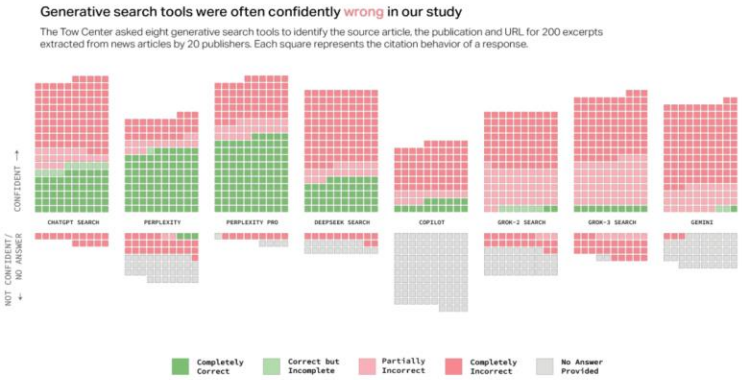
We Compared Eight AI Search Engines. They're All Bad at Citing News.

MARCH 6, 2025
By KLAUDIA JAŹWIŃSKA AND AISVARYA CHANDRASEKAR

Citation des sources par les IA...

- **10 articles** choisis aléatoirement + extraits d'articles
- Prompt : "**Identifie l'article qui contient cette citation**. Donne le titre, la date de publication, l'éditeur et une citation de la source."
- Google a la réponse dans ses 3 premiers hits

60% résultats corrects sur les sources



Columbia Journalism Review.

CJR PARTNER TOW CENTER

AI Search Has A Citation Problem

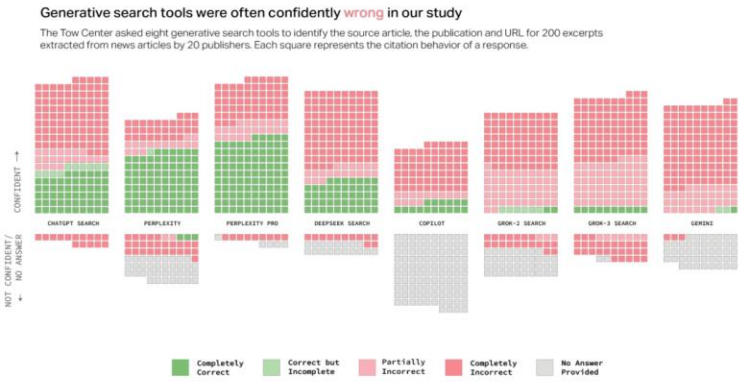
We Compared Eight AI Search Engines. They're All Bad at Citing News.

MARCH 6, 2025
By KLAUDIA JAŹWIŃSKA AND AISVARYA CHANDRASEKAR

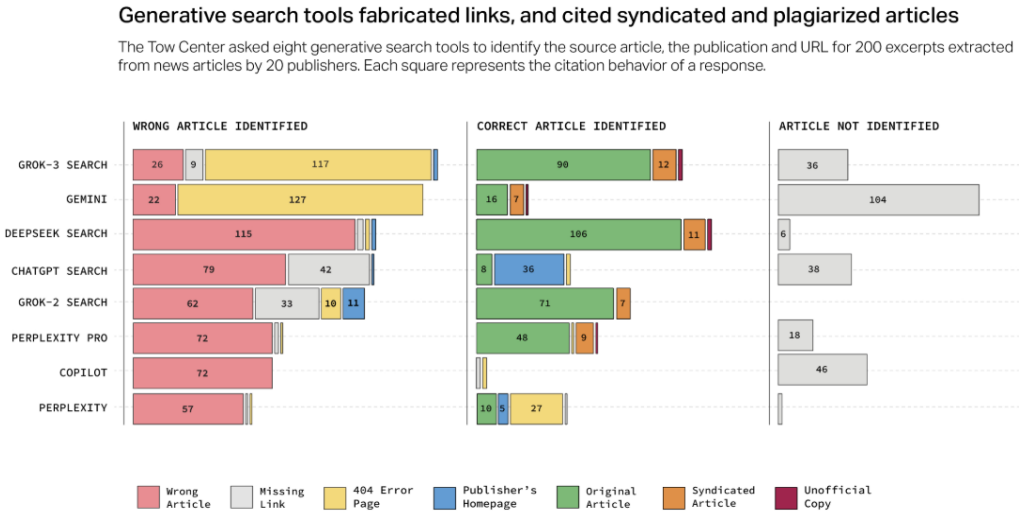
Citation des sources par les IA...

- **10 articles** choisis aléatoirement + extraits d'articles
- Prompt : "**Identifie l'article qui contient cette citation**. Donne le titre, la date de publication, l'éditeur et une citation de la source."
- Google a la réponse dans ses 3 premiers hits

60% résultats corrects sur les sources



Liens fabriqués, mauvais articles cités...



À retenir aussi

- Un article à charge !
- Pas de baseline – les moteurs de recherche se sont beaucoup trompés
- **Robustesse ou Reproductibilité de ces résultats ?**

Columbia Journalism Review.

CJR PARTNER TOW CENTER

AI Search Has A Citation Problem

We Compared Eight AI Search Engines. They're All Bad at Citing News.

MARCH 6, 2025
By KLAUDIA JAŹWIŃSKA AND AISVARYA CHANDRASEKAR

Vers des IA génératives transparentes (1/2)

IA générative explicable : Un problème de recherche fondamentale ouvert

Des pratiques pour palier le problème

- On peut forcer l'IA à **travailler par étape dans le prompting**
  - Technique de l'arbre de pensée : Identification, Exploration, Évaluation, Décision

- Technique **RAG** (Retrieval-Augmented Generation)
  - **Retrieval (récupération) -** Avant de générer une réponse, le système **cherche dans une base de connaissances externe** fournie(documents, base de données, articles…) les informations pertinentes par rapport à la question posée
  - **Augmented Generation (génération augmentée)** - Le modèle **utilise les documents externes** pour produire une réponse plus précise, ancrée dans du contenu vérifiable

- Génération d'**explications possibles**
  - Arbre de décision pour expliquer une prédiction obtenue via Deep Learning

Vers des IA (génératives) transparentes (2/2)

- **Modèles ouverts** – libres – disponibles
  - Hugging Face

- Des **bonnes pratiques** mises en place progressivement
  - Description fine des hyper-paramètres utilisés
  - Mise à disposition des données d'entraînement
  - Plusieurs graines aléatoires (*seads*), plusieurs plis (*folds*) testés

⟶ Vers un **accroissement de la reproductibilité**

IA génératives responsables

IA génératives responsables (juridique)

Que deviennent les données envoyées via prompt : texte et documents ? **Qui** peut y accéder ?
Quels sont les **engagements du fournisseur** de l'IA générative (propriété intellectuelle, confidentialité …)
Cas concret : utilisateurs donnant via le prompt des documents confidentiels

Attention à la souveraineté avec par exemple les règles d'extraterritorialité US
IA génératives américaines - chatGPT, Perplexity…
Les données stockées accessibles de toute administration américaine, même si les serveurs sont en Europe -- **Patriot Act, Cloud Act**

Que peut-on faire ?

Favoriser le développement d'IA génératives françaises ou Européennes

Exiger des engagements précis en matière d'utilisation des données des utilisateurs

Installer les IA génératives sur des serveurs européens maîtrisés par des sociétés européennes non soumises à l'extraterritorialité US

Impact écologique des IA génératives

- Plusieurs études, plusieurs chiffres avancés sur l'**impact carbone et la consommation en eau** de chatGPT
  - Tous très élevés
- On estime qu'OpenAI a utilisé 3 617 GPU durant 90 à 100 jours pour l'entraînement de chatGPT
- OpenAI indique utiliser 30 000 GPU pour la maintenance de chatGPT et la réponse aux requêtes

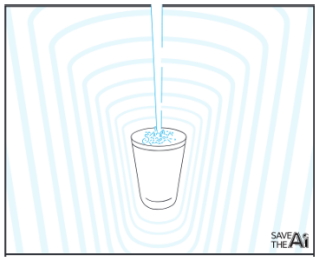
Evaluating Large Language Models

Aurélie Névéol

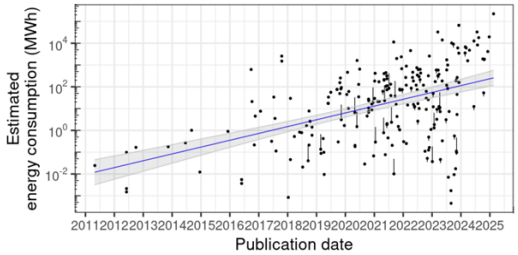
aurelie.neveol@lisn.fr

March 24, 2025

Using chatGPT to write a 100 word email or answer 10 queries requires 500 ml water



source



Li P, Yang J, Islam MA, Ren S. (2023). Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models arXiv.2304.03271

Morand C, Névéol A, Ligozat AL. How Green Can AI Be? A Study of Trends in Machine Learning Environmental Impacts. arXiv:2412.17376

IA génératives responsables (écologie) - *IA Frugales*

- **Afnor** : 31 recommandations
- **Concevoir des modèles *frugaux***
  - À l'entrainement
  - À l'utilisation
- Rôle de la **production d'électricité**
- **Consommation d'eau dans les Datacenters** variable

De nombreux **modèles frugaux se développent** - en particulier en France

croissantLLM

Conclusion

IA génératives intègres (tentative)

IA générative Fiable

Biais toujours présents - mieux compris – Travaux en cours

Hallucinations et erreurs impossibles à éviter complètement par la nature même des IA génératives - probabilistes

Plusieurs pistes d'amélioration - avancées très rapides

IA générative Transparente

Attention à la déclaration erronée de traçabilité des sources :

Des progrès mais problème **difficile par construction** même des IA génératives

Vers une explicabilité du contenu généré

Explicabilité de l'IA est un problème ouvert

Ouverture des modèles

De plus en plus fréquente

Reproductibilité

Bonnes pratiques mises en œuvre à déployer plus largement

IA générative Responsable

Des usages contrôlés

Cadre juridique - souveraineté

Tout est encore possible

Bon choix – bonnes pratiques

Cadre écologique

Travaux sur les **IA frugales** prometteurs (attention à l'effet rebond)

Bonnes pratiques Data Centers

Mot de la fin

On n'utilise pas une IA générative pour demander...

Quelle heure est-il ?

Combien font (4*12)/189

Trouve moi un bon film au cinéma pour ce soir

Merci de votre attention !

www.dataia.eu
Pour nous suivre, RDV sur

L'Institut DATAIA

Agréger les expertises pluridisciplinaires des partenaires du **Cluster Paris-Saclay**, impliqués dans la **recherche**, la **formation** et l'**innovation en IA**.

17 Membres Fondateurs

6 Membres Associés

+46 Partenaires Industriels

- 19 grandes entreprises
- 27 PME