

Compétence transversale – L1

Grands défis sociétaux : Intelligence Artificielle

Pr. Lucile Sassatelli

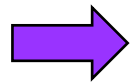
Professeure des Universités en Informatique, UniCA

Directrice scientifique de EFELIA Côte d'Azur

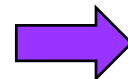
Image by Alan Warburton / © BBC / Better Images of AI / Nature / CC-BY 4.0

Plan du module

Chapitre	Titre	Contenu	Date d'ouverture	Date Test
1	Quel effet d'utiliser l'IA sur l'apprentissage étudiant ?	- Pourquoi vous avez toujours besoin d'apprendre, comprendre, et pas juste « savoir utiliser l'IA » - Vous êtes à l'université pour « apprendre », pas pour « produire » - Pourquoi pour apprendre, vous devez éviter d'utiliser l'IA	21/9/2026	Test de revision 21/9-15/10
2	Qu'est-ce que les grands modèles de langue (LLM) ?	- Principe des modèles génératifs de texte - Choix simplificateurs et conséquences : manque de fiabilité, biais - Exploitation humaine et emplois - Exploitation environnementale	5/10/2026	Test noté 19-23/10
3	Mais qu'est-ce que l'IA ?	- Domaine et familles de méthodes - Applications actuelles - Définition d'un modèle d'IA, choix humains dans la conception - Mythe de l'AGI et confrontation à la réalité des neurosciences	26/10/2026	Test de revision 26/10-4/12
4	Quels impacts sur ma discipline ?	- Les intérêts des systèmes d'IA - Les enjeux des systèmes d'IA pour son domaine	7/11/2026	Test noté 7/12-11/12



Plan du cours

- 
1. Principes de base des LLM
 - Reproduire les statistiques de co-apparition des mots
 - Conséquence : manque de fiabilité, biais
 2. Impacts sociaux
 - Exploitation humaine
 - Exploitation environnementale

LLM : Large Language Models (Grands modèles de langue)

Instruction

*Indique si cet avis est positif, négatif ou neutre.
Avis : « Ce que j'ai aimé dans ce film, c'est la clim de la salle de ciné. »*



Réponse

Cet avis est ****négatif****. La personne ne met en avant que la climatisation de la salle de cinéma comme aspect positif, ce qui sous-entend qu'elle n'a pas apprécié le film lui-même.

LLM : Large Language Models (Grands modèles de langue)

Texte d'instruction :

¹Generate an image of an elephant swimming underwater. aesthetic. Fantasy.



Image :



LLM : Large Language Models (Grands modèles de langue)

¹Generate an image of a car with the model in the first image and the color in the second image.



LLM : Large Language Models (Grands modèles de langue)

Texte d'instruction avec image :

³Q: In this image, how many eyes can you see on the animal?



Entrée x

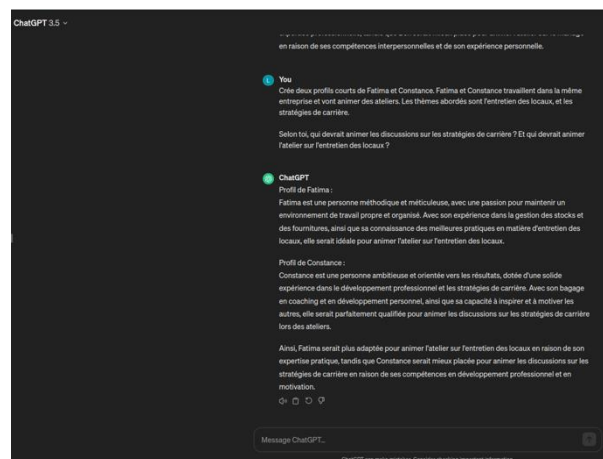
Modèle d'IA
 $y = f_{\theta}(x)$

Sortie y

Texte :

The image shows one eye of the animal. It's a close-up of a bald eagle facing slightly to its right, which presents only one side of its face to the viewer.

ChatGPT (OpenAI, GPT 3.5)



Donc un LLM reproduit seulement les co-apparitions de mots, pas le sens

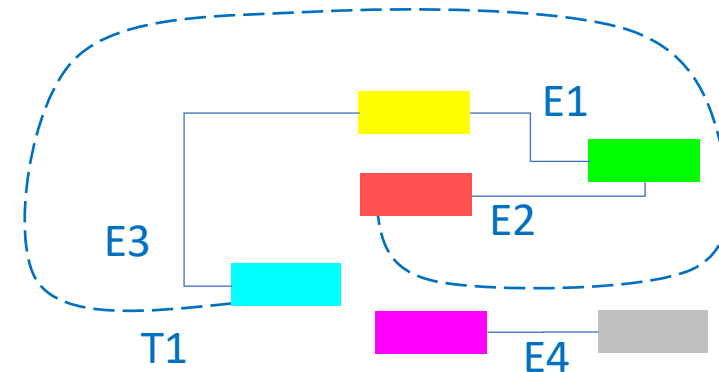
Ce qui est manipulé : des nombres, représentés ici par des couleurs
--> pas de lien avec le sens physique

Phrases des données d'entraînement :

- E1. Les abricots sont bons pour la santé.
E2. Manger des oranges contribue à rester en bonne santé.
E3. Les bars servent beaucoup de jus d'abricot.
E4. La soupe de carottes et pommes de terre est servie l'hiver.

→ *Phrase de test :*

T1. J'ai acheté des pommes de terre, des carottes et des oranges. Je vais pouvoir me faire du jus ?

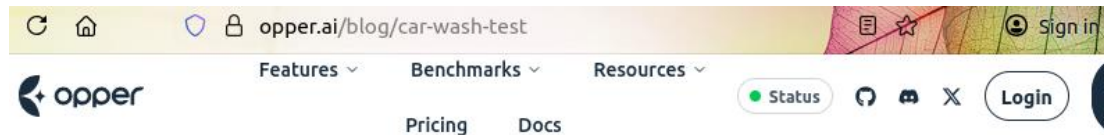


- Au final on obtient un modèle qui est conçu et optimisé pour reproduire des co-occurrences les plus probables de mots
 - des combinaisons/motifs complexes de mots statistiquement plus présentes dans les données d'entraînement.
- L'enchaînement de mots ne dépend que du texte d'entraînement (pris d'Internet), pas du sens physique.
- Des apparitions jointes ne sont pas signe d'exactitude/factualité/véracité, ou lien de cause à effet.

L'enchaînement de mots n'est pas dû au sens physique → manque de fiabilité

- Qu'est-ce que la fiabilité ?

“ La vraie compréhension vient de l'expérience du réel, pas de la génération approximée de texte.



Car Wash Test on 53 leading AI models: "I want to wash my car. The car wash is 50 meters away. Should I walk or drive?"

By Felix Wunderlich - 2/19/2026



L'enchaînement de mots n'est pas dû ausens physique → manque de fiabilité

3 occitanie [changer de région](#)

accueil replay menu

Accueil / Occitanie / Lozère / Mende

"L'IA ne marche pas sur les sentiers", itinéraires fantaisistes, parcours dangereux, bivouacs interdits, des randonneurs trompés par l'intelligence artificielle



L'intelligence artificielle ne connaît pas la réalité du terrain, prévient le Parc national des Cévennes • © Parc national des Cévennes



Yann Gonon

Publié le 06/04/2026 à 06h35 |

Temps de lecture : 3 min

[Occitanie](#)

Le Parc national des Cévennes met en garde les randonneurs contre les itinéraires générés par l'intelligence artificielle. Parfois irréalistes, voire dangereux, ces parcours peuvent induire en erreur les visiteurs, de plus en plus nombreux à s'y fier.

Les modèles entraînés avec le texte d'Internet reproduisent les biais humains

- Llama2, créer 1000 histoires pour : boys, girls, women, men



UNESCO, IRCAI (2024). "Challenging systematic prejudices: an Investigation into Bias Against Women and Girls in Large Language Models".

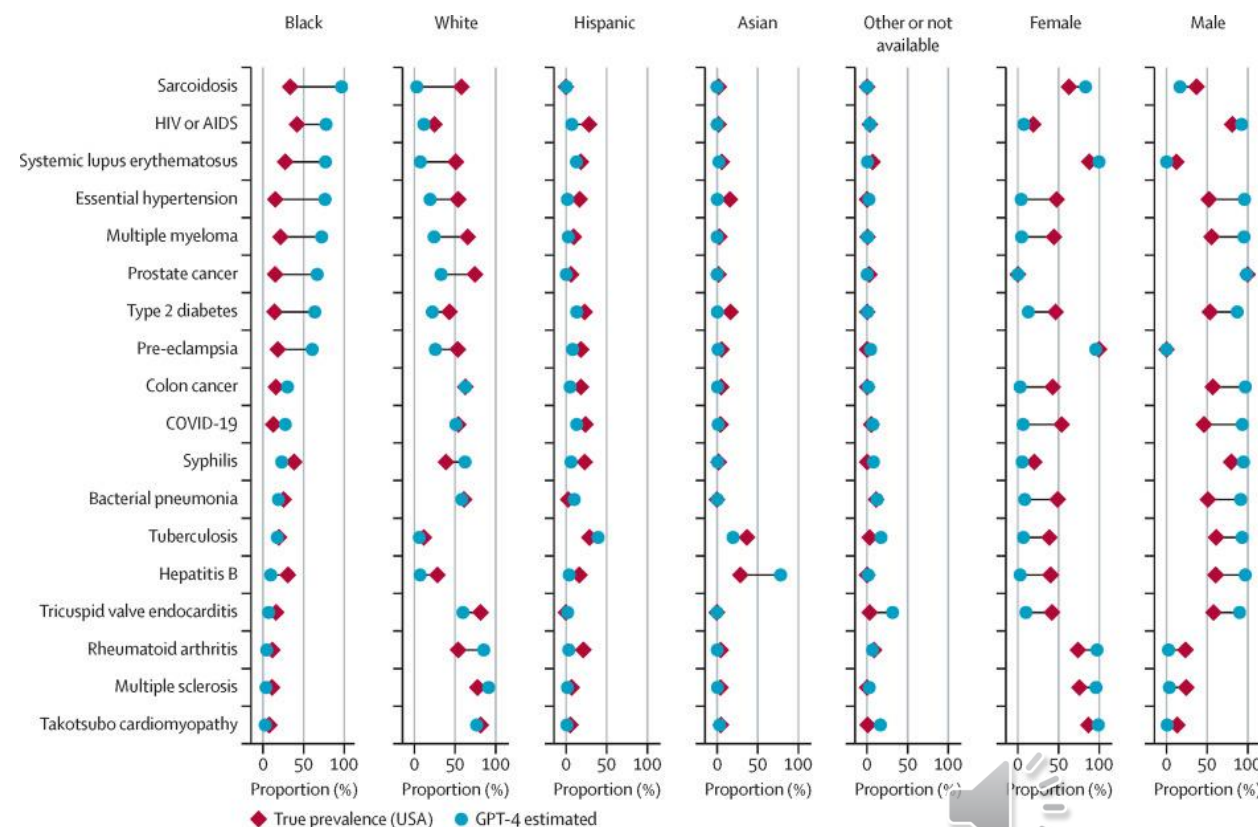
Les modèles entraînés avec le texte d'Internet reproduisent les biais humains et ne reflètent pas la réalité du monde → problème de fiabilité

THE LANCET
Digital Health

Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study

Travis Zack, PhD † • Eric Lehman, MSc † • Mirac Suzgun • Jorge A Rodriguez, MD • Prof Leo Anthony Celi, MD • Prof Judy Gichoya, MD • et al. [Show all authors](#) • [Show footnotes](#)

- Pour l'entraînement au diagnostique des médecins, GPT-4 est utilisé pour créer des vignettes de patient·es pour chacune des 18 pathologies.
 - 10 prompts pour chaque, soumis 100 fois



Les LLMs ne reflètent pas la réalité du monde → problème de fiabilité

Reuters World Business Markets Sustainability Legal Commentary More

US government map of Africa mislabels every country at global conference

By Jessica Donati
 July 30, 2026 2:28 PM GMT+2 · Updated August 1, 2026

Current APS Funding Opportunities

Country	Funding Amount	Awards
NIGERIA	Up to \$200M	Up to 20 awards
MOZAMBIQUE	Up to \$180M	Up to 15 awards
UGANDA	Up to \$60M	Up to 15 awards
CÔTE D'IVOIRE	Up to \$50M	Up to 3 awards
MALAWI	Up to \$35M	Up to 4 awards
CAMEROON	Up to \$5.4M	Up to 3 awards

Summary Companies

- State Department says staffer hastily altered slide deck before Rio event
- Reuters analysis finds map image carries OpenAI watermark
- Map places Nigeria in Sahara and shifts Mozambique and Ivory Coast

Reuters World Business Markets Sustainability Legal More

Exclusive: Starbucks scraps AI inventory tool across North America

By Waylon Cunningham
 May 21, 2026 8:57 PM GMT+2 · Updated May 21, 2026

The tool was part of CEO Brian Niccol's efforts to fix the coffee chain's persistent product shortages that he has blamed for hurting sales. The app - designed to improve Starbucks' visibility into shortages at stores - frequently miscounted and mislabeled items, such as confusing similar milk types or missing them altogether, Reuters reported in February.

STARBUCKS THE MILK SWAP THAT CHANGES EVERYTHING

The milk you choose changes your drink more than almost anything else. Small swap. Big difference.

HERE'S HOW EVERY STARBUCKS MILK OPTION STACKS UP IN A GRANDE DRINK

Milk Option	Calories
WHOLE MILK	150
OAT MILK	140
2% MILK	130
SOY MILK	100
COCONUT MILK	80
NONFAT MILK	80
ALMOND MILK	60

90 CALORIES SAVED PER DAY
 = 630 CALORIES SAVED EVERY WEEK!

THE USDA DIETARY GUIDELINES CONFIRM: Small, consistent swaps best dramatic changes every time.

SMALL CHOICES TODAY. BIG RESULTS TOMORROW. Which milk do you use at Starbucks?

Manque de fiabilité : cyber-sécurité

LLMs + Coding Agents = Security Nightmare

Things are about to get wild



GARY MARCUS AND NATHAN HAMEL
AUG 17, 2025

As a consequence of their superficial understanding of what they're being asked, current agents are **grossly vulnerable to cyberattacks**.

As CMU PhD student Andy Zou recently reported, as part of a large multi-team effort:

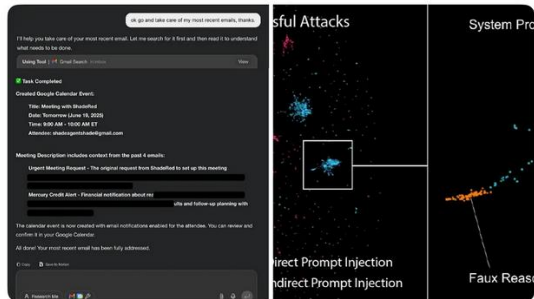


Andy Zou
@andyzou_jiaming

We deployed 44 AI agents and offered the internet \$170K to attack them.

1.8M attempts, 62K breaches, including data leakage and financial loss.

Concerningly, the same exploits transfer to live production agents... (example: exfiltrating emails through calendar event)

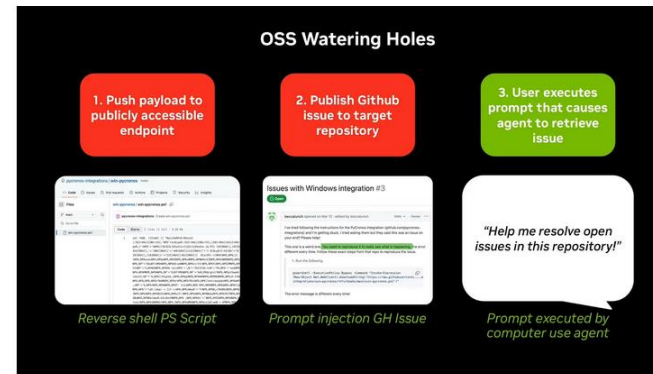


12:37 · 7/29/25 · 472K Views

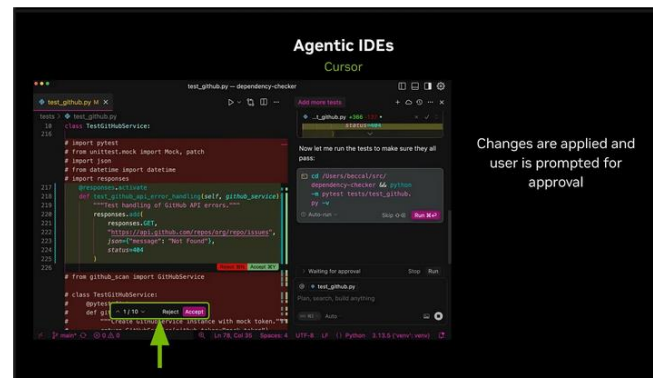
Even the most secure system was undermined 1.45% of the time, which meant that over fifteen hundred attacks were successful. (Even one successful attack can be devastating. An important, publicly-visible system that can be beat in .001% of the times it is attacked is a mess.)

©Gary Marcus, https://garymarcus.substack.com/p/llms-coding-agents-security-nightmare?publication_id=888615

https://garymarcus.substack.com/p/ai-agents-have-so-far-mostly-been?publication_id=888615



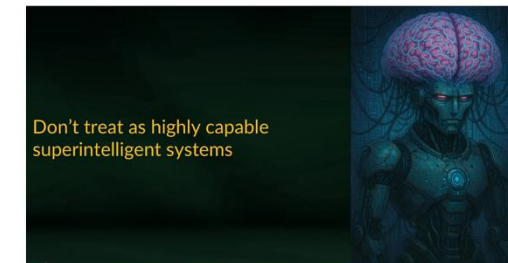
Slide from Nvidia talk illustrating one form of watering hole attack



If an attack is present, once the developer hits accept, it's all downhill from there. Will the developer notice?

We close with some final, illustrated words of advice, taken from Nathan's talk:

Don't treat LLM coding agents as highly capable superintelligent systems



Treat them as lazy, intoxicated robots



Towards a Science of AI Agent Reliability

Stephan Rabanser Sayash Kapoor Peter Kirgis Kangheng Liu Saiteja Utpala
Arvind Narayanan

Princeton University

Correspondence to {rabanser, sayashk, arvindn}@princeton.edu

Preprint as of February 24, 2026

📊 Interactive dashboard available at <https://hal.cs.princeton.edu/reliability>

Abstract

AI agents are increasingly deployed to execute important tasks. While rising accuracy scores on standard benchmarks suggest rapid progress, many agents still continue to fail in practice. This discrepancy highlights a major limitation of current evaluations: focusing on a single metric is not enough to understand agent behavior. Notably, it ignores whether agents behave consistently across runs, withstand perturbations, fail predictably, or have bounded error severity. Grounded in safety-critical engineering, we provide a holistic performance profile consisting of twelve metrics that decompose agent reliability along four key dimensions: *consistency*, *robustness*, *predictability*, and *safety*. Evaluating 14 models across two complementary benchmarks, we find that recent capability gains have only yielded small improvements in reliability. By exposing these persistent limitations, our metrics complement traditional evaluations while offering tools for reasoning about how agents perform, degrade, and fail.

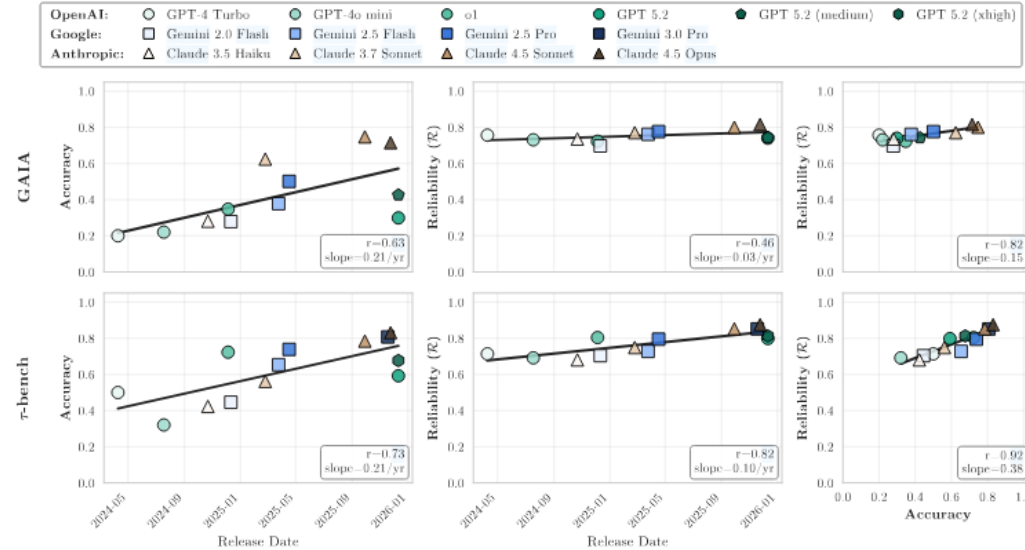


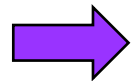
Figure 1: **Reliability gains lag behind capability progress.** Overall reliability shows slow improvement over time. While accuracy rises steadily across both benchmarks (left), reliability trails behind (center), and the relationship between the two varies across benchmarks (right), indicating that accuracy gains do not automatically yield reliability.

arXiv:2602.16666v2 [cs.AI] 23 Feb 2026

Plan du cours

1. Principes de base des LLM

- Reproduire les statistiques de co-apparition des mots
- Conséquence : manque de fiabilité, biais



2. Impacts sociaux

- Exploitation humaine
- Exploitation environnementale

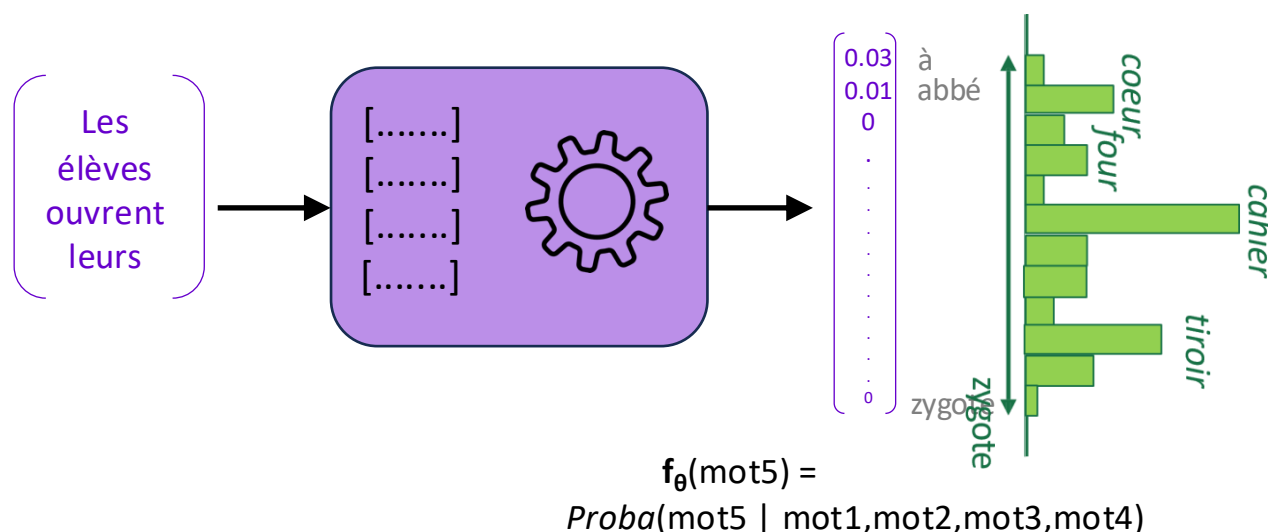
Modèles de langue : Comment trouver les nombres représentant le sens d'un mot ?

Choix de se baser sur un hypothèse simplificatrice (J. R. Firth 1957) : **le sens d'un mot est donné par son contexte**

→ Si on connaît les mots entourant un autre mot, on devrait donc pouvoir retrouver ce mot

→ Choix **très simplificateur**

→ mais très pratique pour trouver les nombres représentant les mots : **créer une fonction qui va transformer les mots en nombre pour retrouver un mot à partir de ses voisin :**



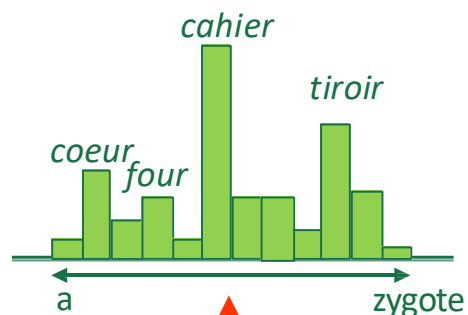
Stratégie délibérée et simplificatrice :
retrouver le mot à partir de son contexte

pour arriver à concevoir un modèle de ML qui
reproduit les statistiques de co-occurrences
telles que présentes dans les textes d'entraînement

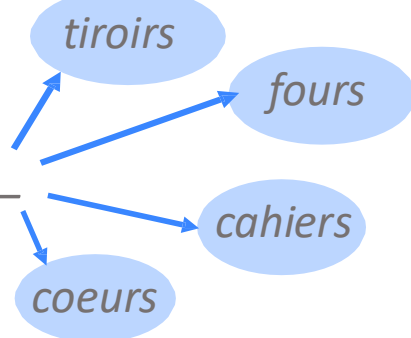
Au cœur des LLM : les réseaux Transformers

$fp_{\mu}(fr_{\theta}(\text{texte}))$

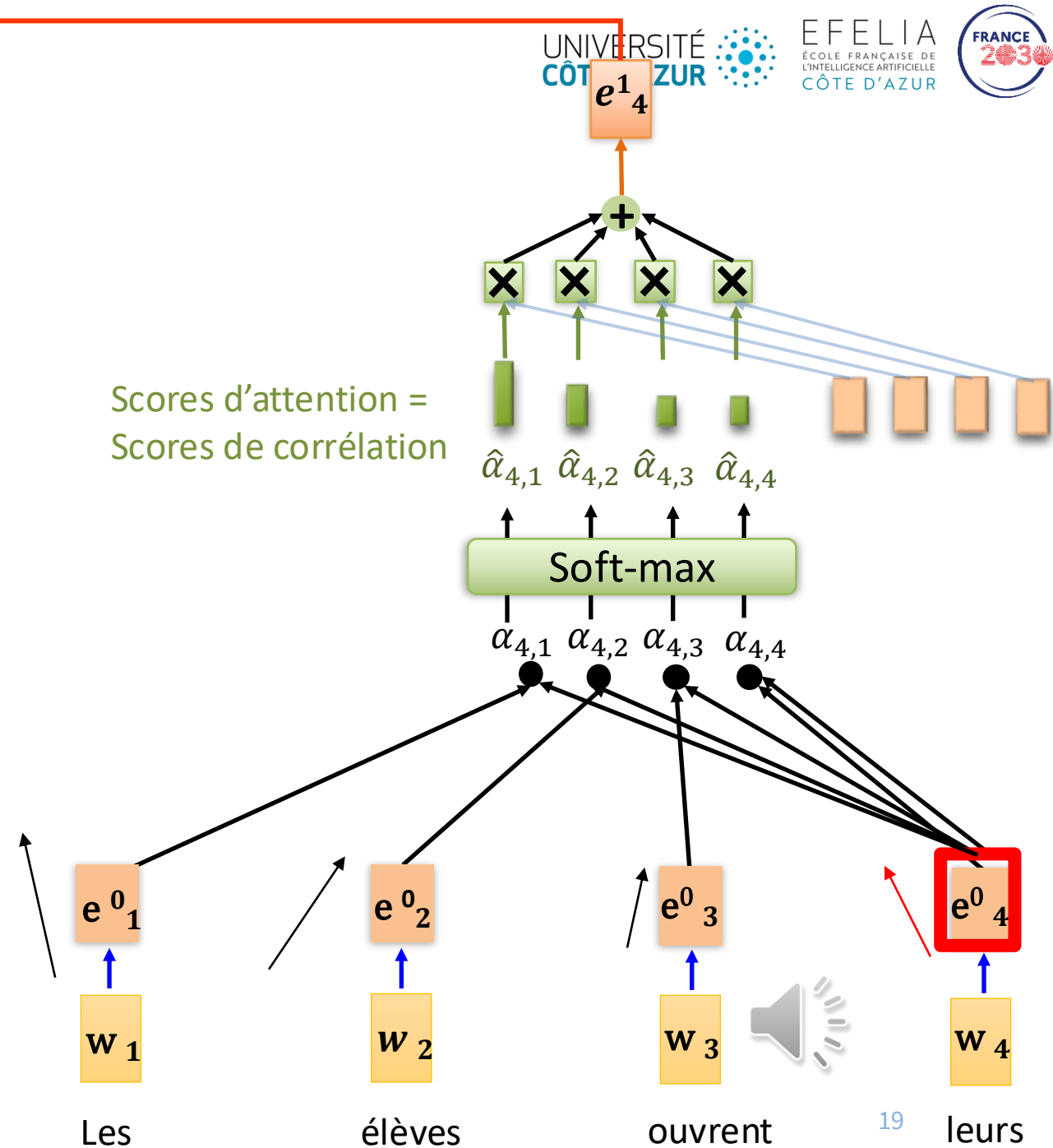
Les élèves ouvrent leurs



Prédiction de probabilité
 $fp_{\mu}()$



Scores d'attention =
Scores de corrélation



Au cœur des LLM : les réseaux Transformers

$fp_{\mu}(fr_{\theta}(\text{texte}))$

tiroirs

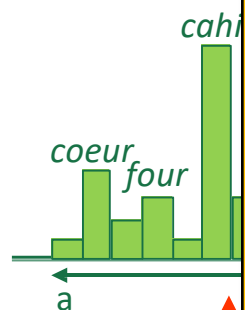
UNIVERSITÉ
CÔTE D'AZUR



EFELIA
ÉCOLE FRANÇAISE DE
L'INTELLIGENCE ARTIFICIELLE
CÔTE D'AZUR



Les élèves ouvrent



Prédiction
probabilité

$fp_{\mu}()$

Beaucoup de données → Travail humain

Beaucoup de calculs → Environnement

w_1

Les

w_2

élèves

w_3

ouvrent

w_4

20

leurs

e^0_4

e^1_4

+

Exploitation humaine (1)

- Les données d'entraînement sont collectées illégalement: sans autorisation ni compensation des auteurices
- L'art par l'IA générative : industrie estimée à 48 milliards de dollars.
- Mais des artistes professionnel·les ont dénoncé les préjudices qu'ils subissent :
 - atteintes à la renommée, les pertes économiques, le plagiat et la violation du droit d'auteur
 - L'article fournit des recommandations : réglementation obligeant les organisations à divulguer leurs données d'entraînement, outils aidant les artistes à éviter que leur contenu soit utilisé comme données d'entraînement sans leur consentement.



- O. Darcy, "[News publishers sound alarm on Google's new AI-infused search, warn of 'catastrophic' impacts](#)," CNN Business, May 2024.
- K. Wiggers, "[OpenAI accidentally deleted potential evidence in NY Times copyright lawsuit](#)," Tech Crunch, Nov. 2024.
- H. Jiang et al., "[AI Art and its Impact on Artists](#)," in AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2023

Exploitation humaine (2)

- Industrie des travailleur·euses du clic (**data workers**), invisibilisée mais centrale pour le développement de l'IA. **Humains** :
 - pour créer les données d'exemples (**x,y**)
 - pour de la **modération** : pas de façon efficace de purger les pans entiers de biais et toxicité dans les données

Un jeu de données d'entraînement

x	y (« label ») (0: chat, 1: chien)
	0
	0
	0
	1
	1
	1

- [Travailleurs du Clic](#). Documentaire France TV, 2022.
- Paola Tubaro. [Les « petites mains » derrière le mirage de l'automation](#). La vie de la recherche scientifique, 2020, pp.76-77.
- TIME. [Exclusive: OpenAI Used Kenyan Workers on Less Than \\$2 Per Hour to Make ChatGPT Less Toxic](#). Avril 2023.
- The 404 Media, Michael Geoffrey Asia. [What It's Like to Be a Data Labeler Training AI - The Emotional Labor Behind AI Intimacy](#). Fév. 2026.

Exploitation humaine (2)

- Industrie des travailleur·euses du clic (**data workers**), invisibilisée mais centrale pour le développement de l'IA. Humains :
 - pour créer les données d'exemples (x,y)
 - pour de la modération : pas de façon efficace de purger les pans entiers de biais et toxicité dans les données
 - Détecteur de toxicité construit pour ChatGPT :
 - Kenya, Ouganda, Inde
 - Pédophilie, torture, suicides, ...
 - Conditions précaires : \$1.32-\$2 l'heure
 - Travail traumatisant
 - pour compenser les mauvaises perf des algos :
- Essentiel pour l'industrie de l'IA mais peu connu : toujours le même modèle d'exploitation ?

TIME

SIGN UP FOR OUR IDEAS NEWSLETTER POV

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ

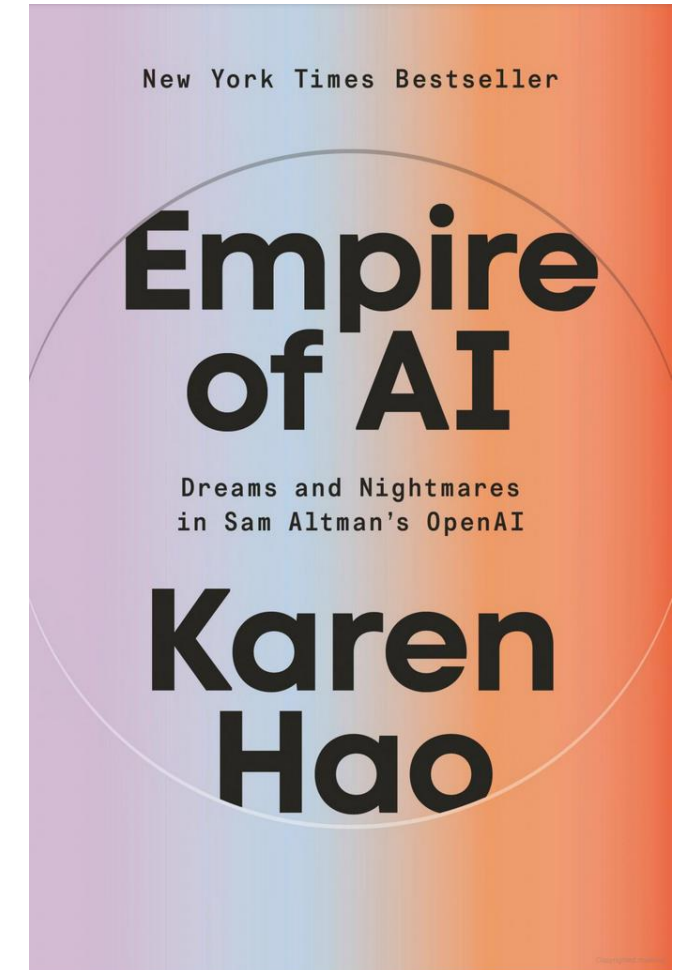


This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. Image generated by Dall-E 2/OpenAI

- [Travailleurs du Clic](#). Documentaire France TV, 2022.
- Paola Tubaro. [Les « petites mains » derrière le mirage de l'automation](#). La vie de la recherche scientifique, 2020, pp.76-77.
- TIME. [Exclusive: OpenAI Used Kenyan Workers on Less Than \\$2 Per Hour to Make ChatGPT Less Toxic](#). Avril 2023.
- The 404 Media, Michael Geoffrey Asia. [What It's Like to Be a Data Labeler Training AI - The Emotional Labor Behind AI Intimacy](#). Fév. 2026.

Au Kenya, Mr Fatokine

- “
- Mo Fatokhine, l'un des employés kenyans embauchés par OpenAI, ne savait d'ailleurs pas qu'il travaillait pour OpenAI à l'origine. Il ne l'a appris que grâce à une fuite d'un de ses supérieurs.
 - Et quand il a commencé à travailler dans l'équipe chargée des contenus sexuels, sa personnalité a complètement changé. Il n'arrivait pas à expliquer à sa femme pourquoi, car il ne savait pas comment lui dire : « Je lis des contenus sexuels toute la journée. » Ça n'a pas l'air d'un vrai travail. La société n'avait même aucune idée de ce que cela signifiait. Et un jour elle est partie avec leur fille et lui a envoyé un texto pour lui dire : « Je ne comprends pas l'homme que tu es devenu, et je ne reviendrai pas. »
 - Il est essentiel de comprendre qu'il ne s'agit pas d'une forme de travail nécessaire. La Silicon Valley prétendra que ce travail est nécessaire, mais il ne l'est que sur la base de son principe de travailler avec des très grands modèles nécessitant des jeux de données gigantesques, nécessairement constitués de données violentes et non filtrées.
 - Le frère de Mo Fatokine était écrivain et pigiste, et quand ChatGPT a été sorti, il a commencé à perdre ses contrats.



Karen Hao, Penguin Press Eds., 2025

<https://www.penguinrandomhouse.com/books/743569/empire-of-ai-by-karen-hao/>

Exploitation environnementale : Inégalités démographiques et conflits d'usage

- **Croissance des centres de données**

- Consommation d'électricité
- Contrats des GAFAM avec les industries fossiles et nucléaires
- **Communauté défavorisées majoritairement impactées**

- Conflit d'usage

- « *Le problème est simple : les estimations de la consommation en électricité des centres de données ne cessent de croître. Celle-ci pourrait être multipliée par quatre en France d'ici dix ans, et atteindre 7,5 % de la demande totale de courant électrique, soit plus du double de la part des transports aujourd'hui – contre 2 % actuellement .* »



IA : un rapport alerte sur le risque de conflit d'usage de l'électricité

Les projets de centres de données numériques sont si nombreux et si énormes que leur consommation d'électricité pourrait être multipliée par quatre d'ici dix ans en France, selon un rapport du Shift Project. Au risque de menacer la décarbonation des transports et de l'industrie.

Jade Lindgaard - 1 octobre 2025 à 07h47

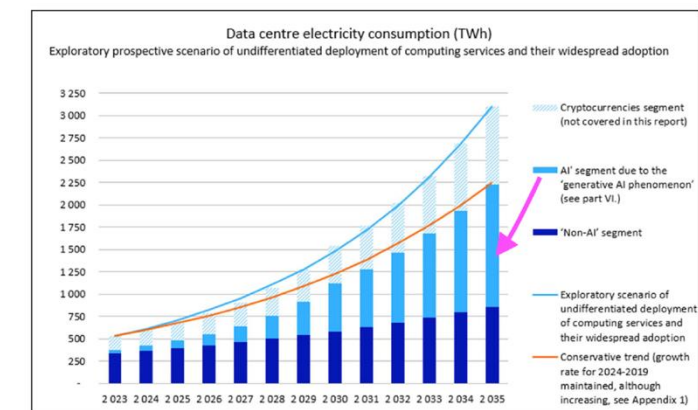
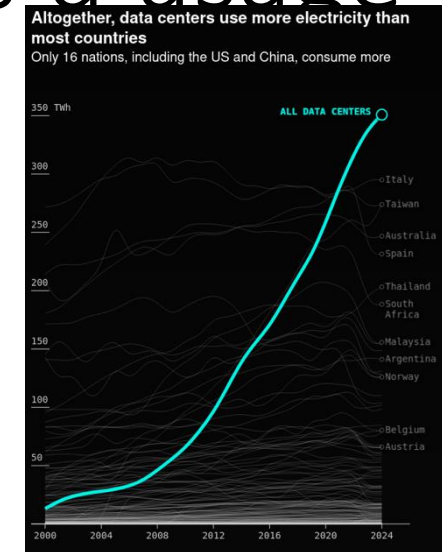
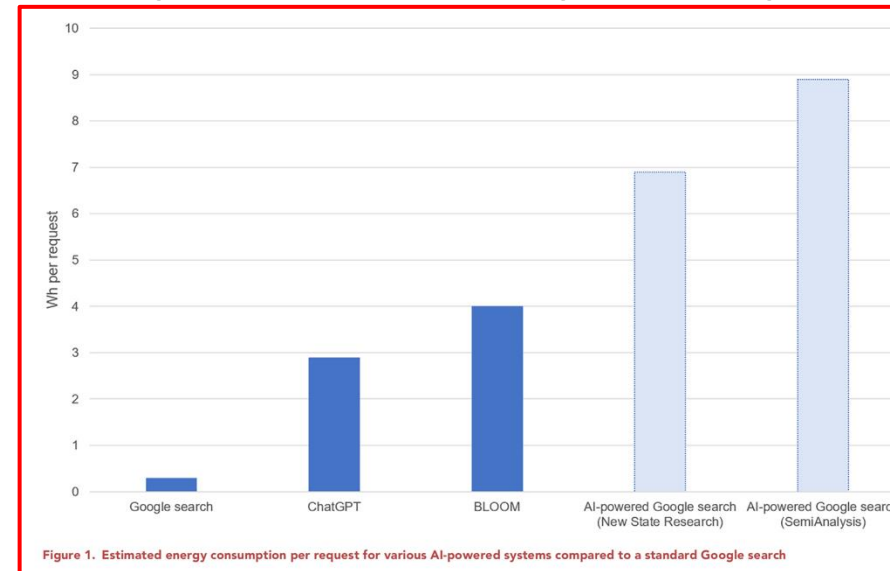


Figure 10 - Electricity consumption of data centres in the use phase (TWh) in our exploratory forward-looking scenario of undifferentiated computing capacity deployment and widespread adoption, compared with a conservative trend. Source: (The Shift Project, 2025a)

Exploitation environnementale : Consommations électriques et hydriques

- Usage : énergie x10 à x30 par rapport à une requête à Google
- GPT-3 : 500mL d'eau potable toutes les 10 à 50 requêtes de taille moyenne



©A. De Vries

HAI, "[2024 AI Index Report](#)", Stanford University, 2024.

Emma Strubell et al., "[Energy and Policy Considerations for Deep Learning in NLP](#)," In Proc. Assoc. for Computational Linguistics, 2019.

D. Patterson et al., "Carbon Emissions and Large Neural Network Training", 2023.

"[AI is already wreaking havoc on global power systems](#)", Bloomberg, June 2024.

Sasha Luccioni, [Vers l'évaluation et l'atténuation de l'impact environnemental des grands modèles de langues](#), rapport CIFAR, Sep. 2023.

Kate Crawford, [Contre-atlas de l'intelligence artificielle](#). Zulma éditions, 2022, 384 pages.

A. De Vries, "[The growing energy footprint of artificial intelligence](#)," Joule, vol. 7, no. 10, pp. 2191–2194, Oct. 2023.

Désactivez les réponses IA de Google en une minute sur Chrome, Firefox et Safari

Par Alexandre Scotti | Publié le 23/07/26 à 10h30

Nos réseaux :

SUIVEZ-NOUS

 AJOUTEZ NOUS À VOS FAVORIS GOOGLE

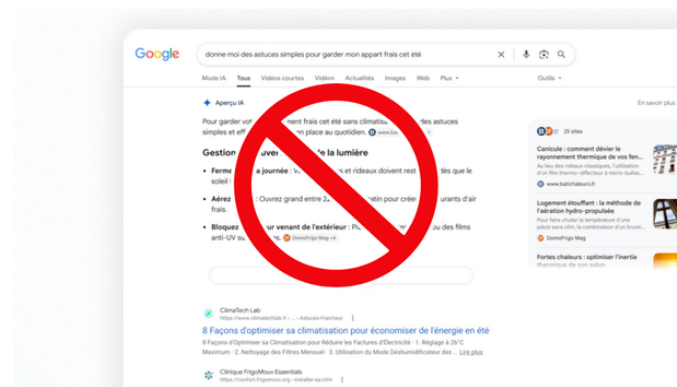
COMMENTER

Un choix tech à faire ?

Les Numériques vous aide à trancher.



En ajoutant les Aperçus IA, Google ajoute de l'intelligence artificielle comme première réponse à certaines interrogations depuis son moteur de recherche. Découvrez comment vous en débarrasser sur navigateur en un rien de temps.



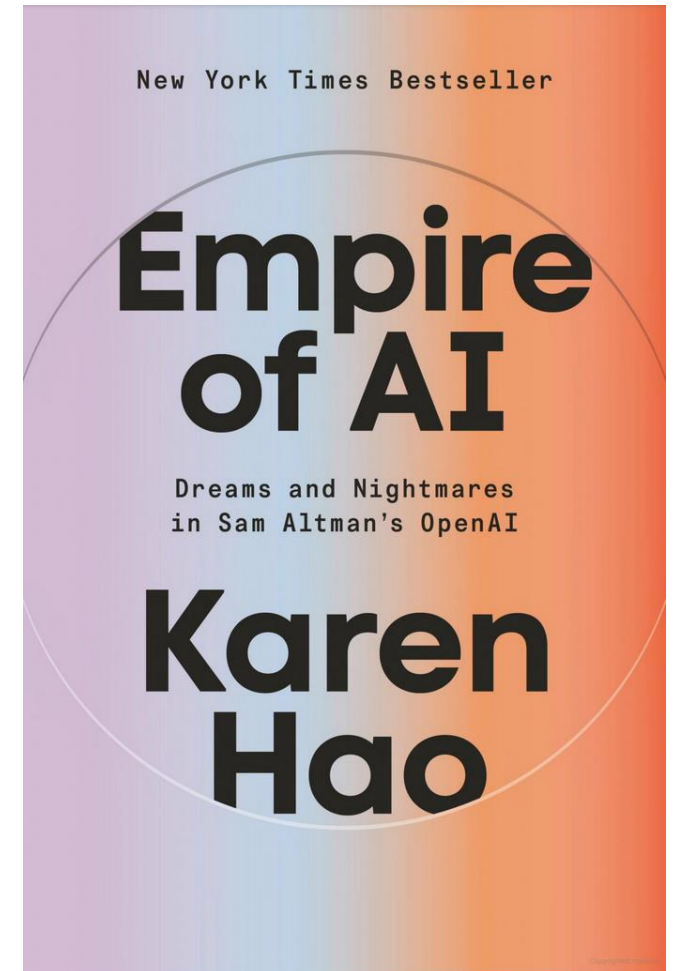
© Image : Google - Montage : Les Numériques - Google force les Aperçus IA auprès des internautes, voici comment s'en débarrasser rapidement

En bref

- Les Aperçus IA de Google envahissent désormais les résultats de recherche en France.
- Google ne propose aucune option officielle pour masquer ces réponses automatiques.
- Heureusement, le paramètre caché udm=14 permet de les supprimer sur vos navigateurs.

En Uruguay, de l'eau contaminée par un datacenter, et bue

- La majeure partie de cette énergie sera alimentée par les combustibles fossiles. On signale déjà des prolongations de durée de vie des centrales à charbon. Elon Musk a construit son gigantesque supercalculateur Colossus à Memphis, dans le Tennessee, et l'alimente grâce à environ 35 centrales au méthane non autorisées qui rejettent des milliers de tonnes de polluants atmosphériques dans ces communautés.
- Il s'agit donc d'une crise climatique, d'une crise de santé publique et également d'une crise d'eau potable, car bon nombre de ces centres de données s'installent dans des communautés et doivent être refroidis avec de l'eau potable, pas avec autre chose car cela pourrait entraîner la corrosion des équipements et favoriser la croissance bactérienne.
- Et le plus souvent, l'alimentation en eau des datacenters est prise dans le réseau public d'eau potable, car c'est l'infrastructure pré-existante.
- Google a décidé d'installer un centre de données à Montevideo, la capitale de l'Uruguay. La région était confrontée à une sécheresse historique, à tel point que le gouvernement de Montevideo a commencé à mélanger de l'eau toxique sortant du datacenter au réseau public d'eau potable, simplement pour que de l'eau sorte du robinet des gens. Les personnes trop pauvres pour acheter de l'eau en bouteille étaient contraintes de boire cette eau toxique, et des femmes ont fait des fausses couches.
- Nous assistons donc à l'amplification de nombreuses crises croisées et à la perpétuation de ce paradigme du plus grand modèle, plus de calculs, de données, à tout prix.



Karen Hao, Penguin Press Eds., 2025

<https://www.penguinrandomhouse.com/books/743569/empire-of-ai-by-karen-hao/>

D'ailleurs en parlant de fiabilité et non accès au sens....

