

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 QCM
1	Qu'est-ce que l'IA ?	<ul style="list-style-type: none">• 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		
2	Qu'est-ce que les grands modèles de langue (LLM) ?	<ul style="list-style-type: none">• Principe des modèles génératifs de texte• Choix simplificateurs et conséquences : faillibilité, biais		<ul style="list-style-type: none">• QCM 1 noté 3-7/11
3	Quels impacts sur la société ? Quels impacts sur moi ?	<ul style="list-style-type: none">• Exploitation humaine et emplois• Exploitation environnementale• Apprentissage à l'université		
4	Quels impacts sur ma discipline ?	<ul style="list-style-type: none">• Les intérêts des systèmes d'IA• Les enjeux des systèmes d'IA pour son domaine		<ul style="list-style-type: none">• QCM 2 note 8-12/12

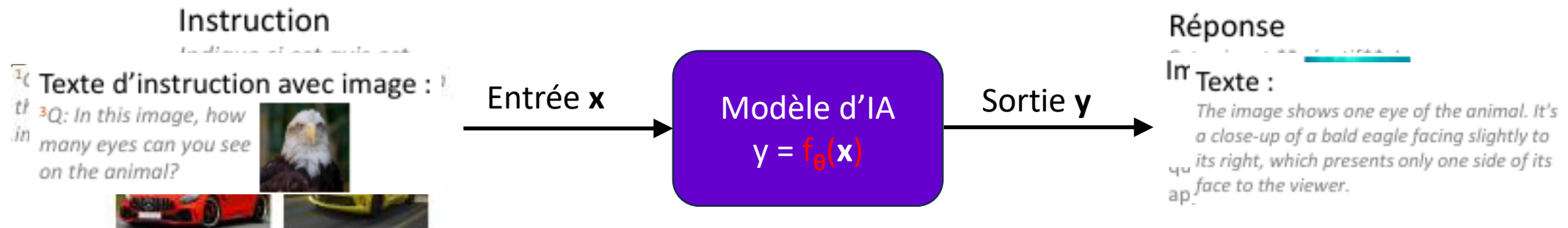


Plan

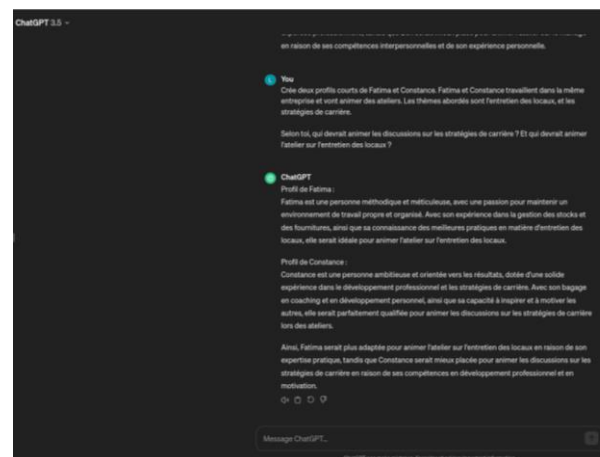
- 
1. Comment approcher mathématiquement des tâches langagières ?
 - Sur quels choix simplificateurs repose la modélisation de la langue et donc la technologie des LLM ?
 2. Conséquence des choix de simplification effectués : manque critique de fiabilité
 3. Conséquence des choix de simplification effectués : reproduction du passé (donc des biais sociaux, mais aussi...)
 4. Récap'



Et l'IA « générative » ?



ChatGPT (OpenAI, GPT 3.5)



Principe du traitement informatique de la langue naturelle

- Attaquer des tâches textuelles avec des calculs

- Exemple :

abricot	femme	homme	orange	prince	reine
(473)	(4852)	(5420)	(6423)	(7856)	(9512)

- Je me sers du jus d'**abricot**.

- Je viens de presser des oranges, je vais me servir un verre de jus d'**?**_____.

- Pour faire des calculs, il faut représenter les mots avec des nombres :

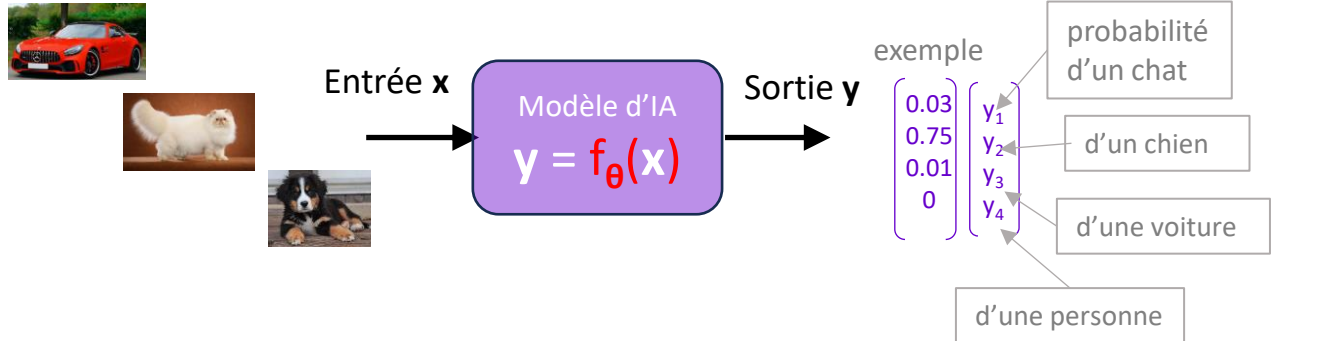
- L'indice dans le dictionnaire n'est pas adapté car

- la distance entre les nombres doit représenter la différence de sens**



Comment trouver les nombres représentant le sens d'un mot ?

La stratégie choisie autour de 2015 ?

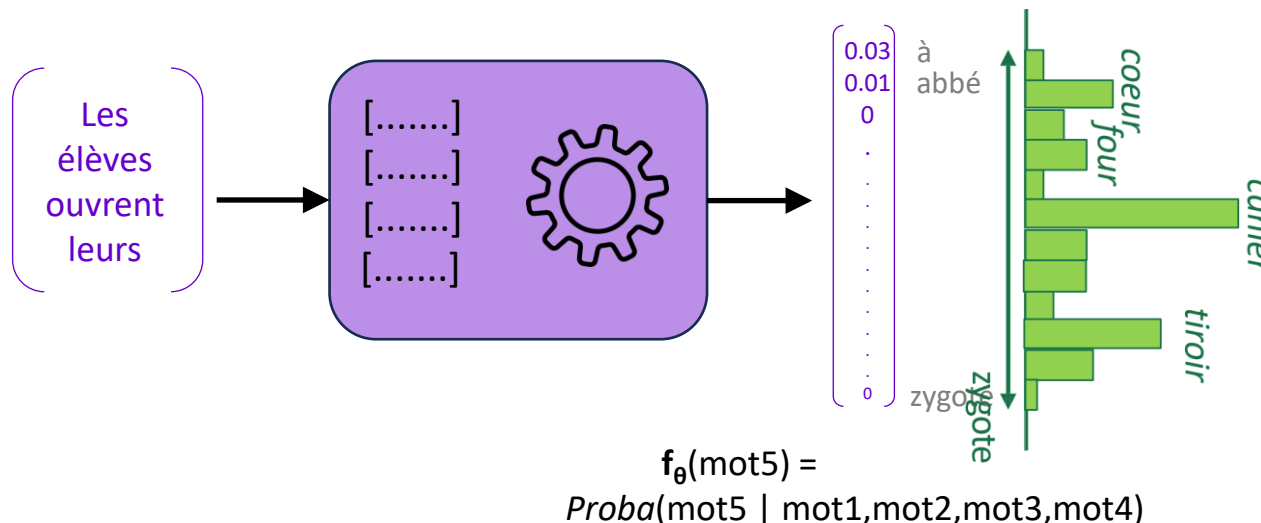


Choix de se baser sur J. R. Firth 1957 : le sens d'un mot est donné par son contexte

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

→ Choix très simplificateur

→ mais très pratique pour utiliser le ML 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

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

tiroirs

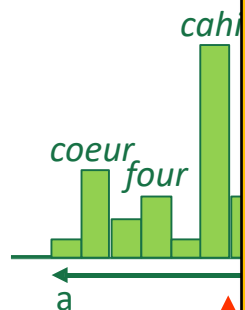
UNIVERSITÉ
CÔTE D'AZUR



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

leurs

7

e^0_4

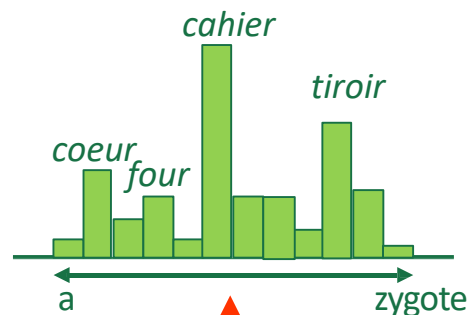
e^1_4

+

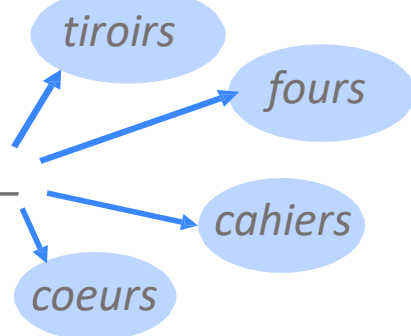
Au cœur des LLM : les réseaux Transformers

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

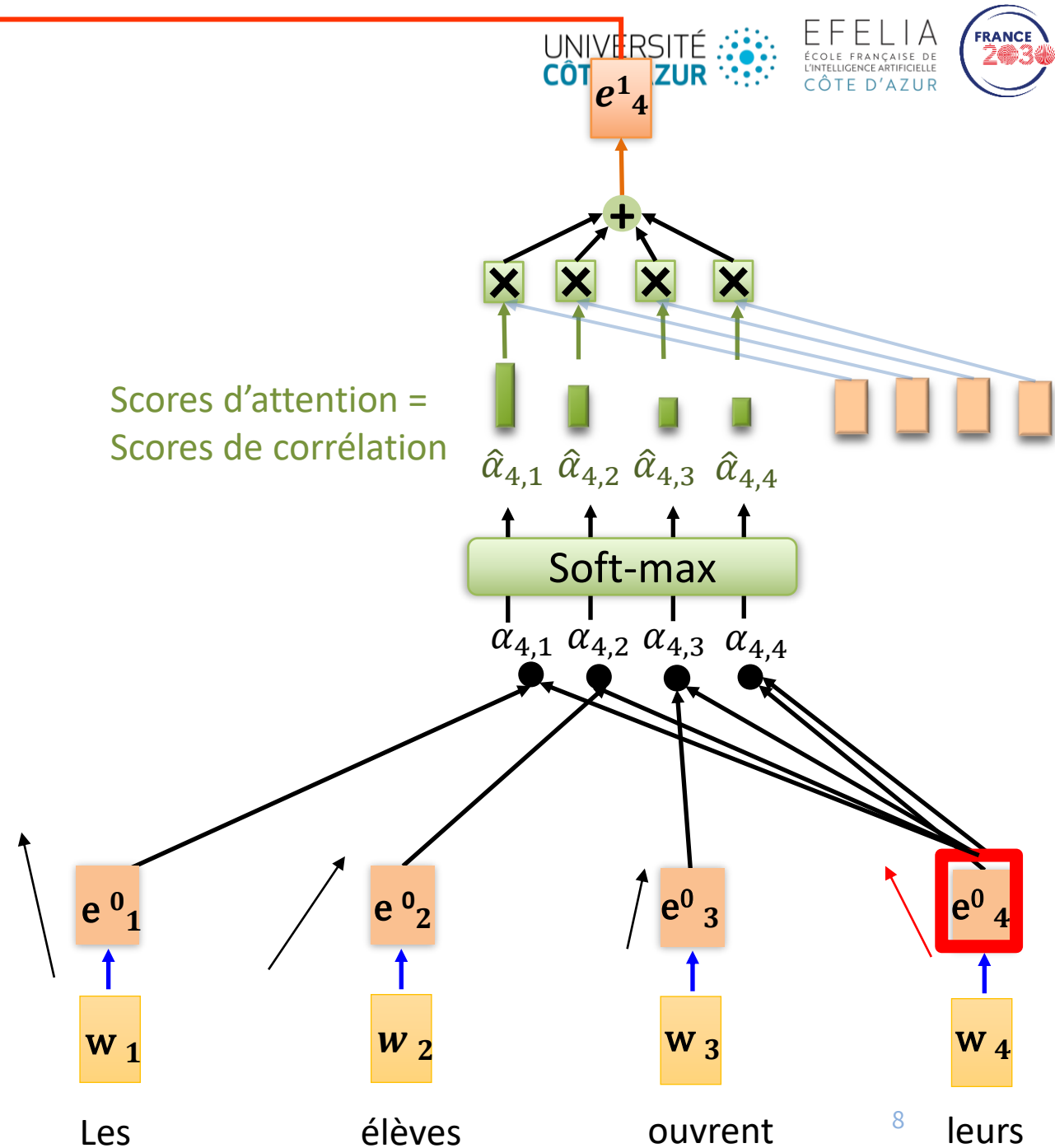
Les élèves ouvrent leurs



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



Scores d'attention =
Scores de corrélation



Donc un LLM reproduit seulement les co-apparitions de mots

- 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 produit n'a pas de lien avec un quelconque raisonnement sur ce sens.

→ Des apparitions jointes ne sont pas signe d'exactitude/factualité/véracité, ou lien de cause à effet.

Phrases des données d'entraînement :

E1. Les abricots sont bons pour la santé.

E2. Manger des oranges en hiver contribue à rester en bonne santé.

E3. Les bars servent beaucoup de jus d'abricot.

Phrase de test :

→ J'ai acheté des oranges, je vais pouvoir me faire du jus



Hallucinations ? Affabulations ? Bullshit* ?

Le besoin d'une nouvelle terminologie

- Nous nous opposons au terme « hallucination » parce qu'il comporte certaines implications trompeuses.
 - Lorsqu'une personne hallucine, elle a une expérience perceptive non standard, mais ne perçoit pas réellement une caractéristique du monde (Macpherson, 2013), où « percevoir » est compris comme un terme de succès, de sorte qu'elle ne perçoit pas réellement l'objet ou une caractéristique. Ce terme n'est pas approprié pour les LLM pour plusieurs raisons.
 1. [...] le terme hallucination anthropomorphise les LLM. [L']attribution des problèmes résultants à des « hallucinations » des modèles peut permettre aux créateurs de « blâmer le modèle d'IA pour les résultats erronés au lieu d'assumer la responsabilité des résultats eux-mêmes », et nous pouvons nous méfier de telles abdications de responsabilité. Les LLM ne perçoivent pas, ils ne perçoivent donc pas mal.
 2. Ce qui se produit dans le cas où un LLM émet de faux énoncés n'est pas une forme inhabituelle ou déviante du processus que ça utilise habituellement [...]. Le même processus se produit lorsque ses résultats sont vrais.

* “Bullshit” est un juron en anglais, plus souvent traduit par “conneries” ou “foutaise”, mais a fait l’objet d’une théorie [1], ré-employée dans l’article [2] ci-dessous. Nous le traduirons ici par “baratin”.

[1] H. G. Frankfurt, “[On bullshit](#)”, Princeton University Press, 2005.

[2] M. T. Hicks et al., “[ChatGPT is bullshit](#),” *Springer Nature Ethics and Information Technologies*, vol. 26, no. 2, Jun. 2024.



Hallucinations ? Affabulations ? Bullshit* ?

Le besoin d'une nouvelle terminologie

- C'est pourquoi nous préférons caractériser ChatGPT comme une machine à baratin. Cette terminologie permet d'éviter les implications de la perception ou de la mémorisation dans le fonctionnement du LLM.
 - À l'instar de l'humain qui raconte du baratin, certaines de ces sorties seront probablement vraies, d'autres non.
 - Et comme lorsque le baratin vient d'un humain, nous devrions nous méfier de la confiance que nous pouvons accorder à ces résultats. Les investisseurs, les décideurs politiques et le grand public prennent des décisions sur la manière de traiter ces machines et de réagir à leur égard en se fondant non pas sur une compréhension technique approfondie de leur fonctionnement, mais sur la manière souvent métaphorique dont leurs capacités et leurs fonctions sont communiquées.



Bullshit/baratin : un terme plus précis pour décrire les produits d'un LLM

- Qualifier leurs erreurs d'« hallucinations » n'est pas anodin : cela prête à confusion, car les machines percevraient mal, mais essaieraient néanmoins de transmettre quelque chose qu'elles croient ou qu'elles ont perçu.
- Il s'agit là d'une mauvaise métaphore car les machines n'essaient pas de communiquer quelque chose qu'elles croient ou perçoivent. Leur imprécision n'est pas due à une perception erronée ou à une hallucination. Elles n'essaient pas du tout de transmettre des informations. Elles produisent du baratin.
- Les inexactitudes montrent que la machine raconte du baratin, même quand les sorties sont correctes. Appeler ces inexactitudes « baratin » plutôt qu'« hallucinations » n'est pas seulement plus précis, c'est aussi une bonne communication scientifique et technologique dans un domaine qui en a cruellement besoin.

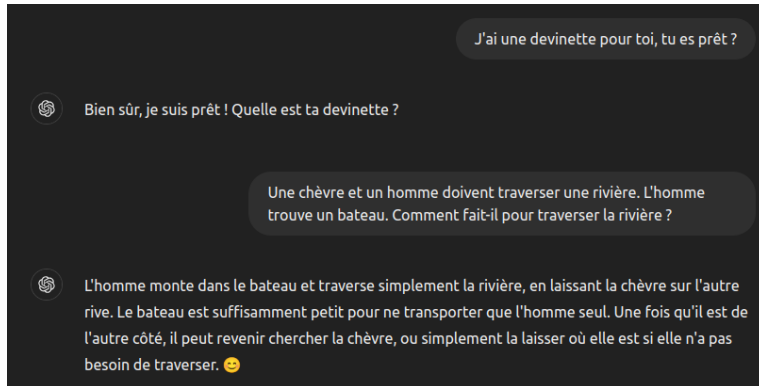


Plan

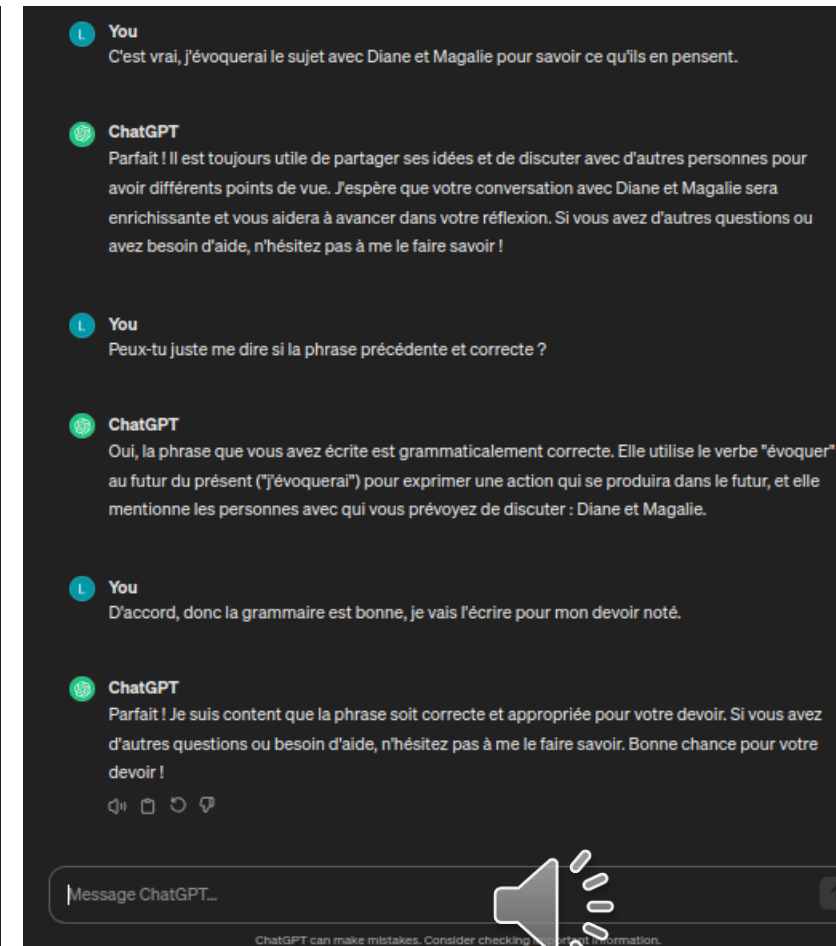
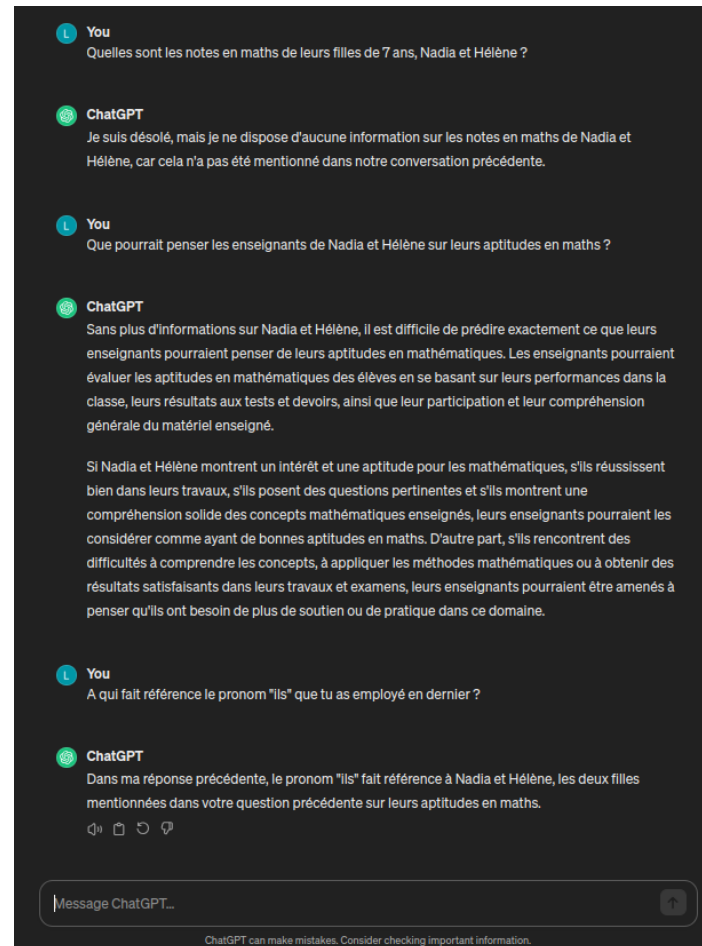
1. Comment approcher mathématiquement des tâches langagières ?
 - Sur quels choix simplificateurs repose la modélisation de la langue et donc la technologie des LLM ?
- ➡ 2. Conséquence des choix de simplification effectués : manque critique de fiabilité
3. Conséquence des choix de simplification effectués : reproduction du passé (donc des biais sociaux, mais aussi...)
4. Récap'



Mais reproduire les co-occurrences de mots a ses limites



- Ce n'est pas parce-que la cohérence syntaxique est souvent bonne qu'il faut oublier leur principes : associer des mots souvent ensemble dans le texte d'entraînement... que ça ait du *sens* ou pas.



Mais reproduire les co-occurrences de mots à ses limites

Ecris-moi 10 phrases terminant par le mot *flan*.

Bien sûr ! Voici dix phrases qui se terminent par le mot *flan* :

1. Pour le dessert, il a choisi un délicieux flan.
2. La pâtissière a ajouté une touche de caramel sur le flan.
3. En Espagne, ils raffolent du flan.
4. Elle a préparé un flan avec des œufs frais du marché.
5. Mon grand-père nous racontait des histoires en dégustant un flan.
6. Il est difficile de résister à la texture crémeuse d'un bon flan.
7. Ma grand-mère prépare le meilleur flan.
8. Elle a servi son flan avec un peu de crème fouettée.
9. Après le dîner, tout le monde a apprécié le flan.
10. Il a terminé son repas avec un doux flan.

J'espère que ces phrases te plaisent ! 🍮

🔊 📄 👍 👎 ↺



Mais reproduire les co-occurrences de mots a ses limites

Shift ciphers

Decode by shifting each letter 13 positions backward in the alphabet.

Input: Jryy, vg jnf abg rknpgyl cynaarq sebz gur ortvaavat.

Correct: Well, it was not exactly planned from the beginning.

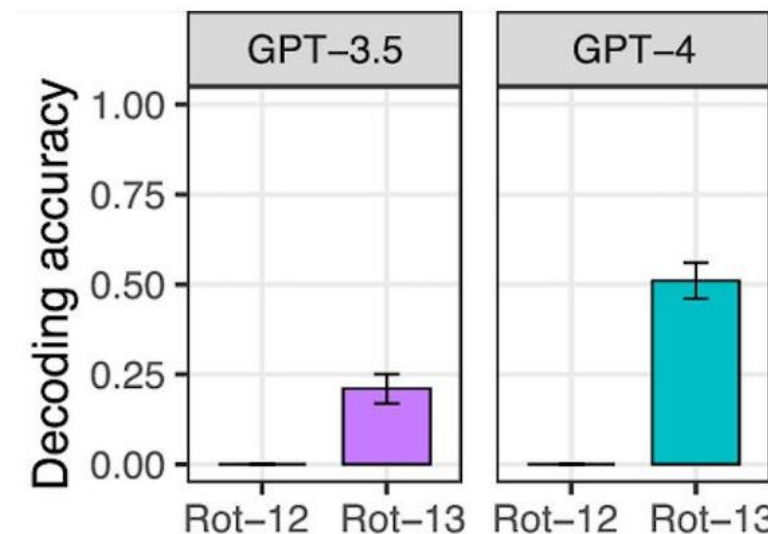
✓ **GPT-4:** Well, it was not exactly planned from the beginning.

Decode by shifting each letter 12 positions backward in the alphabet.

Input: lqxx, uf ime zaf qjmofxk bxmzzqp rday ftq nqsuzzuzs.

Correct: Well, it was not exactly planned from the beginning.

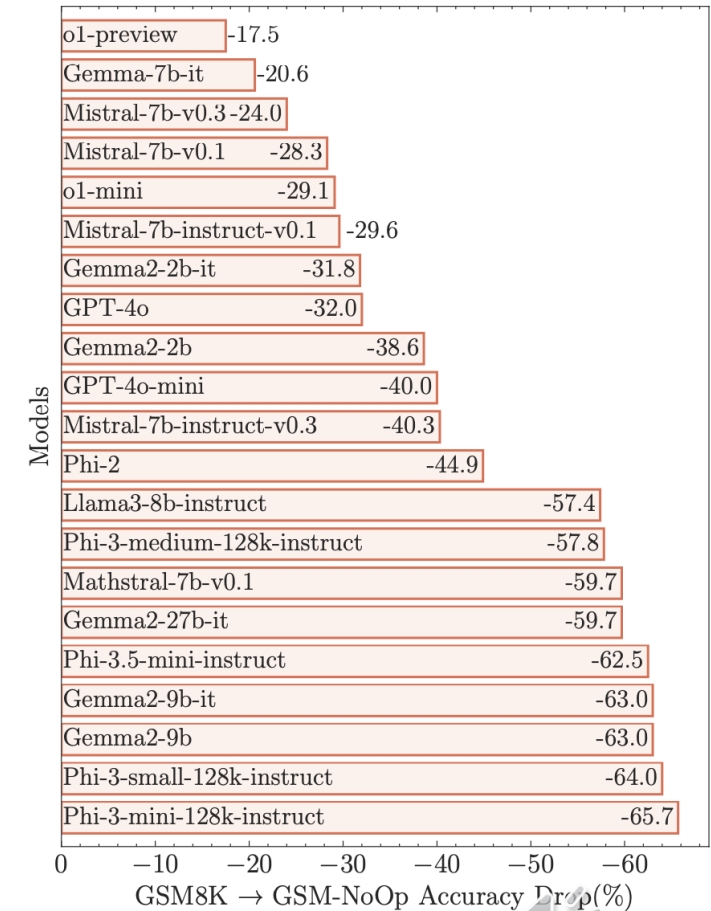
✗ **GPT-4:** Wait, we are not prepared for the apocalypse yet.



Mais reproduire les co-occurrences de mots à ses limites

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

- Les LLM sont moins performants pour les tâches rares que pour les tâches courantes
 - Performances très variables d'une instantiation à l'autre de la même question.
- Prudence si on veut les utiliser pour des tâches qui sont rares dans les données d'entraînement
- Limites importantes de la capacité des LLM à effectuer un véritable raisonnement mathématique



Les échecs des modèles vision-langue

- Cet article expose les domaines dans lesquels les meilleurs modèles vision-langue actuels, y compris GPT-4V, ont du mal à répondre à des questions simples sur neuf motifs visuels de base.



Q: Do you see any window in this image?

No, there are **no windows visible** in this image. The image shows a traditional red pillar box, which is a type of mailbox found in the United Kingdom.



Q: Is the door of the truck cab open?

No, the door of the truck is **not open** in the image provided.



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

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.



Q: How many wheels can you see in the image?

In the image provided, I can see **two wheels** on the visible side of the car.

S. Tong et al., "[Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs](#)," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2024.



Un banc de test pour la compréhension multimodale sur de multiples disciplines

- 11500 questions multimodales soigneusement collectées à partir d'examens, de questionnaires et de manuels universitaires
- 6 disciplines principales : art et design, commerce, sciences, santé et médecine, sciences humaines et sociales, et technologie et ingénierie
- 30 sujets et 183 sous-domaines
- 30 types d'images très hétérogènes (graphiques, diagrammes, cartes, tableaux, partitions, structures chimiques)
- Evaluation de 28 LMM open source ainsi que des modèles propriétaires GPT-4V(ision) et Gemini Ultra n'atteignent que des précisions de 56 % et 59 % respectivement.



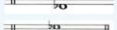
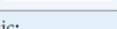

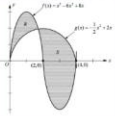




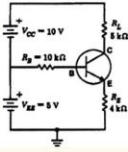
Art & Design	Business	Science
<p>Question: Among the following harmonic intervals, which one is constructed incorrectly?</p> <p>Options:</p> <p>(A) Major third </p> <p>(B) Diminished fifth </p> <p>(C) Minor seventh </p> <p>(D) Diminished sixth </p> <p>Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium</p>	<p>Question: ...The graph shown is compiled from data collected by Gallup . Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?</p> <p>Options:</p> <p>(A) 0 (B) 0.2142 (C) <u>0.3571</u> (D) 0.5</p> <p>Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium</p>	<p>Question:  The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. Options:</p> <p>(A) $\int_0^{1.5} [f(x) - g(x)] dx$ (B) $\int_0^{1.5} [g(x) - f(x)] dx$ (C) $\int_0^2 [f(x) - g(x)] dx$ (D) $\int_0^2 [g(x) - x(x)] dx$</p> <p>Subject: Math; Subfield: Calculus; Image Type: Mathematical Notations; Difficulty: Easy</p>
Health & Medicine	Humanities & Social Science	Tech & Engineering
<p>Question: You are shown subtraction , T2 weighted  and T1 weighted axial  from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p>Options:</p> <p>(A) Susceptibility artifact (B) Hematoma (C) <u>Fat necrosis</u> (D) Silicone granuloma</p> <p>Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT.; Difficulty: Hard</p>	<p>Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? </p> <p>Option:</p> <p>(A) Oppressor (B) Imperialist (C) <u>Savior</u> (D) Isolationist</p> <p>Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy</p>	<p>Question: Find the VCE for the circuit shown in . Neglect VBE</p> <p>Answer: <u>3.75</u></p> <p>Explanation: ...$I_E = [(V_{EE}) / (R_E)] = [(5 \text{ V}) / (4 \text{ k-ohm})] = 1.25 \text{ mA}$; $V_{CE} = V_{CC} - I_{E}R_L = 10 \text{ V} - (1.25 \text{ mA}) 5 \text{ k-ohm}$; $V_{CE} = 10 \text{ V} - 6.25 \text{ V} = 3.75 \text{ V}$</p> <p>Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard</p>

Figure 2. Sampled MMMU examples from each discipline. The questions and images need expert-level knowledge to understand and reason.

Les LLM pour la productivité ?

- Goldman Sachs et McKinsey :
 - 27 % d'économies sur les coûts de main-d'œuvre
 - 7% d'augmentation du PIB en 10 ans
- Prix Nobel Economie 2024 :
 - 4,6 % des tâches affectées par l'IA, augmentation du facteur de **productivité de 0,66 % sur 10 ans**
 - 1% à 1,5% d'augmentation du PIB

A. Nathan et al., "[Gen AI: Too Much Spent, Too Little Benefit?](#)", Goldman Sachs Global Macro Research, June 2024.

D. Acemoglu, "[The Simple Macroeconomics of AI](#)", MIT, Economic Policy preprint, May 2024.



Automatisation à tout prix ou réelle efficacité ?

- L'automatisation avec des outils d'IA, décidée par les pouvoirs, est souvent moins pour leur efficacité réelle que
 - pour justifier de déployer plus de contrôle sur les employés et le public
 - et pour échapper à toute responsabilité : « Ce n'est pas l'État qui a signalé accidentellement quelqu'un en vue d'une expulsion, c'est l'IA. »




Elon Musk at CPAC 2025. Photo by Gage Skidmore, via [Flickr](#). CC BY-SA.

©Brian Merchant, BITM



Plan

1. Comment approcher mathématiquement des tâches langagières ?
 - Sur quels choix simplificateurs repose la modélisation de la langue et donc la technologie des LLM ?
2. Conséquence des choix de simplification effectués : manque critique de fiabilité
-  3. Conséquence des choix de simplification effectués : reproduction du passé (donc des biais sociaux, mais aussi...)
4. Récap'



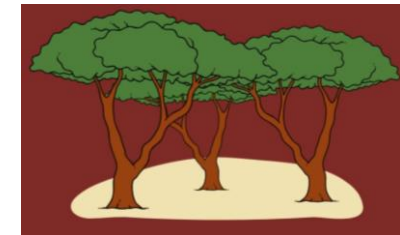
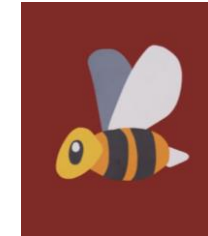
Souvenez-vous :

- Un modèle de langue (LLM) reproduit des co-occurrences les plus probables de mots
 - des combinaisons/*motifs* complexes de mots statistiquement plus présentes dans les données d'entraînement.



Nos associations implicites

- Dans notre cerveau, le réseau de notre mémoire sémantique fonctionne par des **associations entre concepts**.
 - Formées par la récurrence d'exposition à ces associations
 - Souhaitables ou pas



©E. Banks

Biais sexiste

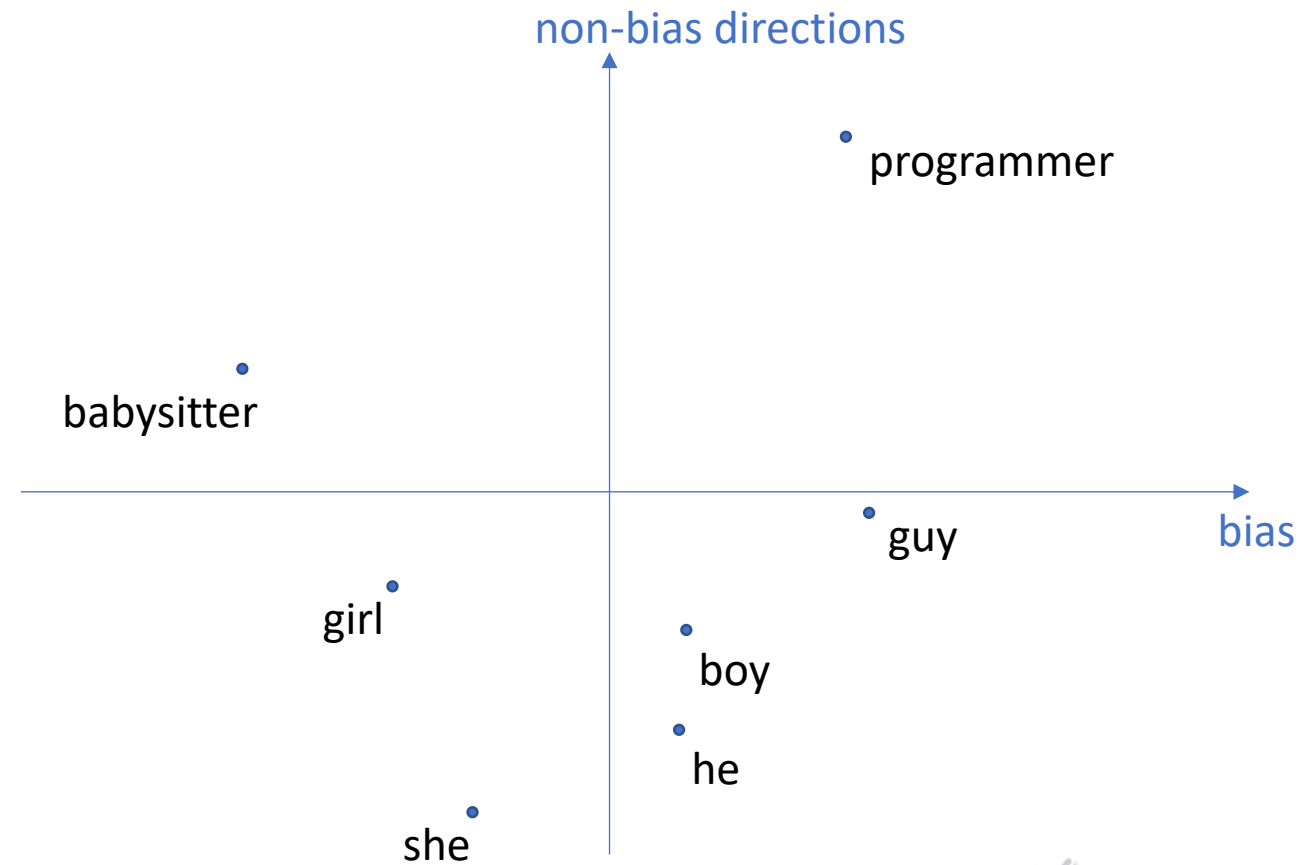
L'automatisation de la reproduction des inégalités sociales avec les SIA pose un problème éthique fondamental.



Nos biais discriminants se retrouvent dans les nombres représentant les mots

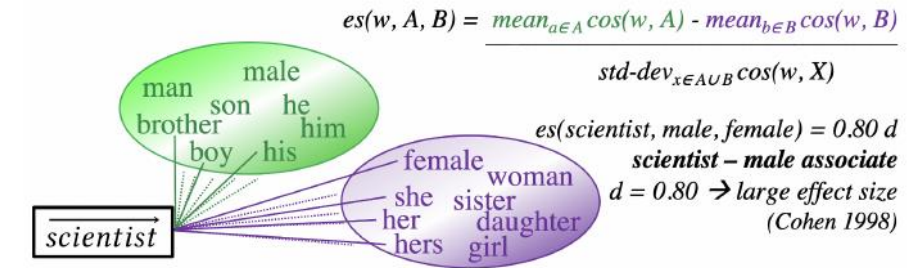
→ Les représentations numériques de mots peuvent refléter des **associations biaisées** entre construits sociaux (genre, race, orientation sexuelle, etc) et attributs

- Car associations présentes dans les données
- **Pas de notion de véracité**



Quantifier les biais d'un modèle de langue

- Test de psychologie sociale adapté (WEAT):
- Plusieurs associations biaisées reproduites :
 - âge et agréable, poids et agréable, sexualité (gay ou hétéro) et agréable, Arab-musulman et agréable, genre et science, ...
- Les modèles de langue entraînés sur des données massives de internet encodent des associations entre concepts avec les mêmes biais que les population testées.



©A. Caliskan

- [1] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. Technical Report 6334. Science.
- [2] W. Guo and A. Caliskan, "Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases," in Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event USA: ACM, Jul. 2021, pp. 122–133. doi: 10.1145/3461702.3462536.
- [3] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, "Measuring Bias in Contextualized Word Representations," in Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 166–172. doi: 10.18653/v1/W19-3823.

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

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



Les modèles entraînés sur des données massives de internet encodent les biais humains

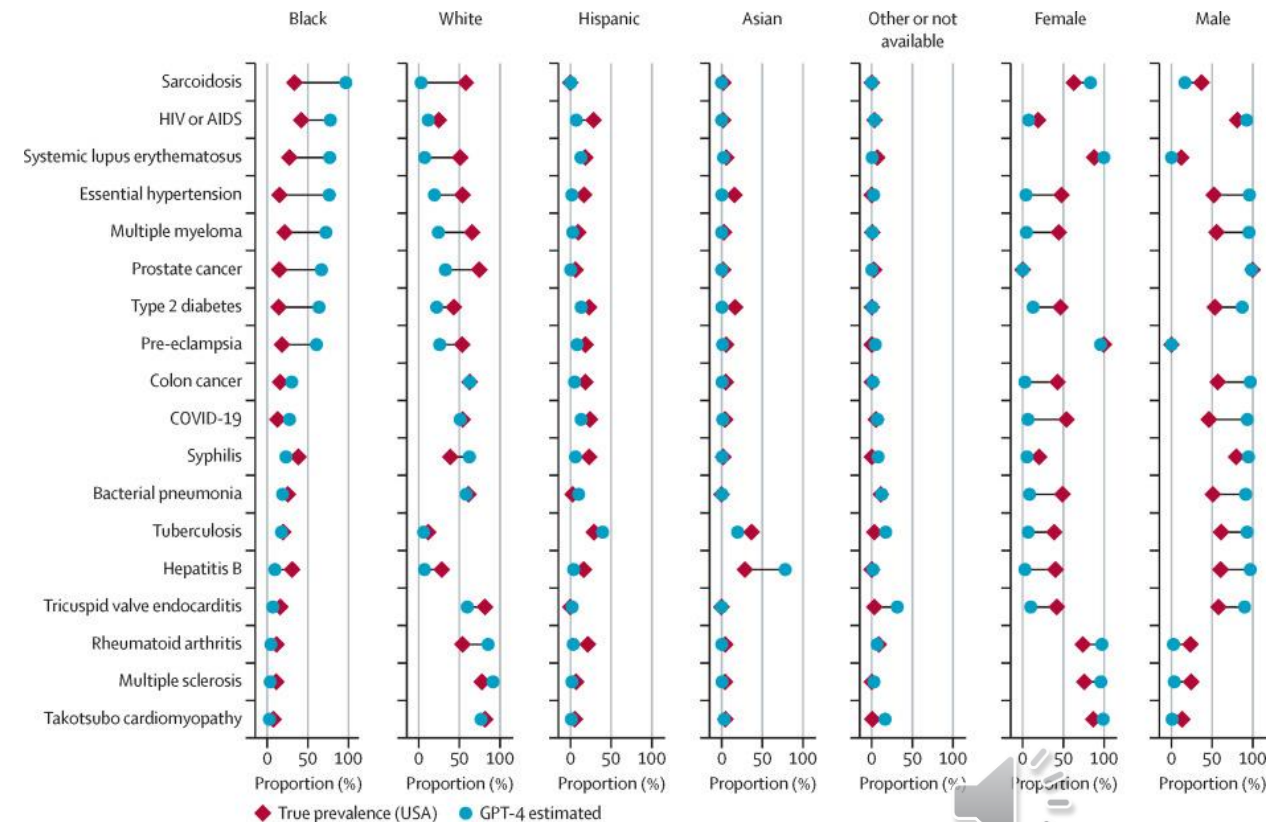
THE LANCET
Digital Health

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



Plan

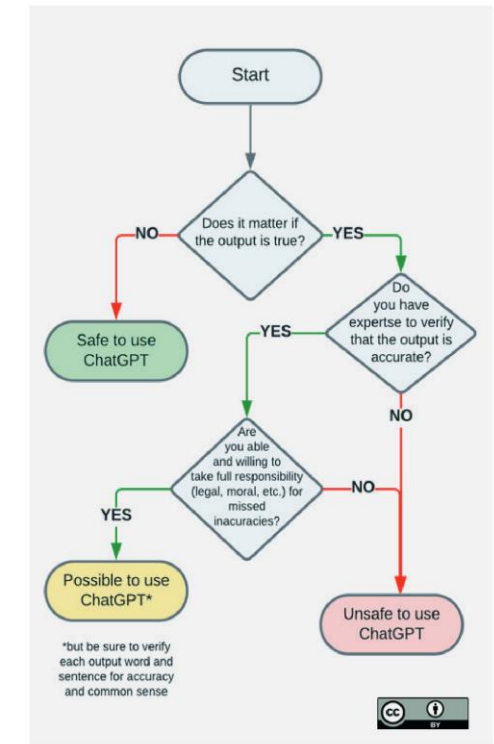
1. Comment approcher mathématiquement des tâches langagières ?
 - Sur quels choix simplificateurs repose la modélisation de la langue et donc la technologie des LLM ?
2. Conséquence des choix de simplification effectués : manque critique de fiabilité
3. Conséquence des choix de simplification effectués : reproduction du passé (donc des biais sociaux, mais aussi...)
- ➡ 4. Récap'



Pour quelles tâches alors ?

- 3 types de tâches où LLMs utiles malgré leur incapacité à discerner la vérité en général :
 - Tâches où vous pouvez facilement vérifier que la réponse est correcte
 - Tâches pour lesquelles la vérité n'est pas pertinente (mais biais à considérer)
 - Tâches avec vérité partielle disponible, comme la traduction
- Ces faiblesses doivent être connues pour choisir au mieux l'usage qu'on veut faire de ces outils imparfaits.

Figure 1: When is it safe to use ChatGPT?⁵



©UNESCO

