

# Limits and challenges of conversational models



• Draw a very biased chatbot. • Add a little bit of steampunk style to this picture.

DALL-E 3 (via ChatGPT, GPT-4, 01/02/2024)

# Training data



- Draw a representation of corpus creation, steampunk style
- DALL-E 3 (via ChatGPT, GPT-4, 2023-11-19)

# Availability of language models

- Recent advances in language modelling have 2 main causes
  - **Bigger and bigger models**
  - **Bigger and bigger training corpora**
- **Some models freely distributed ("open-weights")** – under diverse licences
  - Masked language models: BERT, RoBERTa, ALBERT, CamemBERT, FlauBERT, ELECTRA, DeBERTa, CamemBERTa
  - Generative models: GPT(-1,-2), T5, GPT-Neo, GPT-J, LLaMA(-1,-2), OPT, BLOOM, Pythia, Falcon-40B, Mistral 7B, OLMO
- **The biggest models trained by private actors are rarely freely distributed**
  - Generative models: GPT(-3,-3.5,-4), Claude, GLaM, Gopher, Chinchilla, Megatron, PaLM(-1, -2, -2 large), Falcon-180B, Gemini 1.5

# Training corpora

- Language models are trained on **very large amounts of text**
  - Often extracted from the internet (generally from Common Crawl dumps)
  - Sometimes from other sources (legal?)
  - Quality issues (incorrect information, non-standard style, bias)
  - Legal and ethical issues (copyright, personal data, illicit content...)
- Orders of magnitude: several hundred billion to several trillion words
  - That corresponds to 1 to 5 million times Tolstoy's *War and Peace*
  - Reading such a corpus would take several tens of thousands of years
- **Training corpora are even less often available** than models
  - Open source = open data + open weights = open codebase



Common Crawl



# Classification of data sources

		Is it legal to access this data freely?	Is it legal to use this data to train a model?
White data	Yes	Yes	
Grey data	Yes	Unclear / No	
Black data	No	No	



Dataset  
/bookcorpus



# Which training data are used?

- Sometimes white and grey data sources only
  - OSCAR 2019 for CamemBERT, ROOTS for BLOOM, FineWeb for many recent models
- Often (probably) on white, grey and black data sources
  - Very few recent models – even open-weight ones! – **describe** their training data precisely
    - Example: LLaMA 1
  - Even fewer make it **downloadable**
    - (Approximative) reproduction of LLaMA 1's pretraining corpus: RedPajama (v1)



# Which training data are used?

- Sometimes white and grey data sources only
  - OSCAR 2019 for CamemBERT, ROOTS for BLOOM
- Often (probably) on white, grey and black data sources
  - Very few recent models – even open-weight ones! – **describe** their training data precisely
    - Example: LLaMA 1
  - Even fewer make it **downloadable**
    - (Approximative) reproduction of LLaMA 1's pretraining corpus: RedPajama

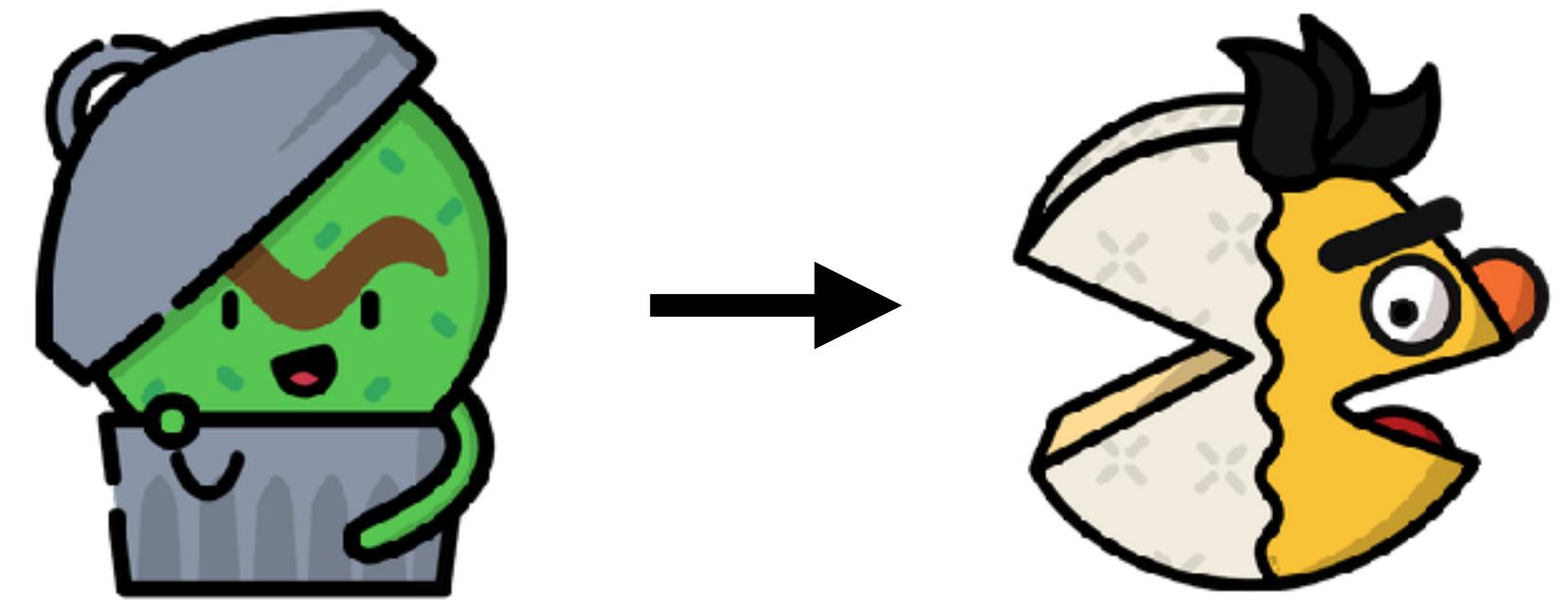
---

Training Data	
LLAMA 1	<i>See Touvron et al. (2023)</i>
LLAMA 2	<i>A new mix of publicly available online data</i>

---

# Quantity, quality, diversity

- The original BERT model was trained on 4B tokens (Wikipedia + Book Corpus ; ~16GB of non-compressed text)
- The main contribution of the **RoBERTa** paper ([Liu et al. 2019](#)): **using more data has a positive impact**
  - They use ~40B tokens, more diverse (160GB of non-compressed text)
- For **CamemBERT** ([Martin et al. 2019](#)), we carried out experiments with the French Wikipedia (1B tokens) and with our internet-based corpus, **OSCAR** (32B tokens)
  - 1B tokens randomly sampled from **OSCAR** is almost enough...
    - (It is **less** than what was thought necessary before)
    - ...and much better than the French Wikipedia:  
**one can use less data provided it is diverse**
  - For least 38 languages, **OSCAR** contains 1B tokens or more



# Quantity, quality, diversity

- High quality data is better
  - Textbook data made it possible to train a code generation model ( $\phi$ -1) that performs very well despite its “small” size (1.3B parameters) ([Gunasekar et al. \(2023; Microsoft; “Textbooks Are All You Need”\)](#))
- But there is a balance between quality and quantity
  - A lot of less strictly filtered data is better than a less data of a higher quality ([Penedo et al. 2023; Falcon team; RefinedWeb](#))
  - Difference between pre-training and fine-tuning

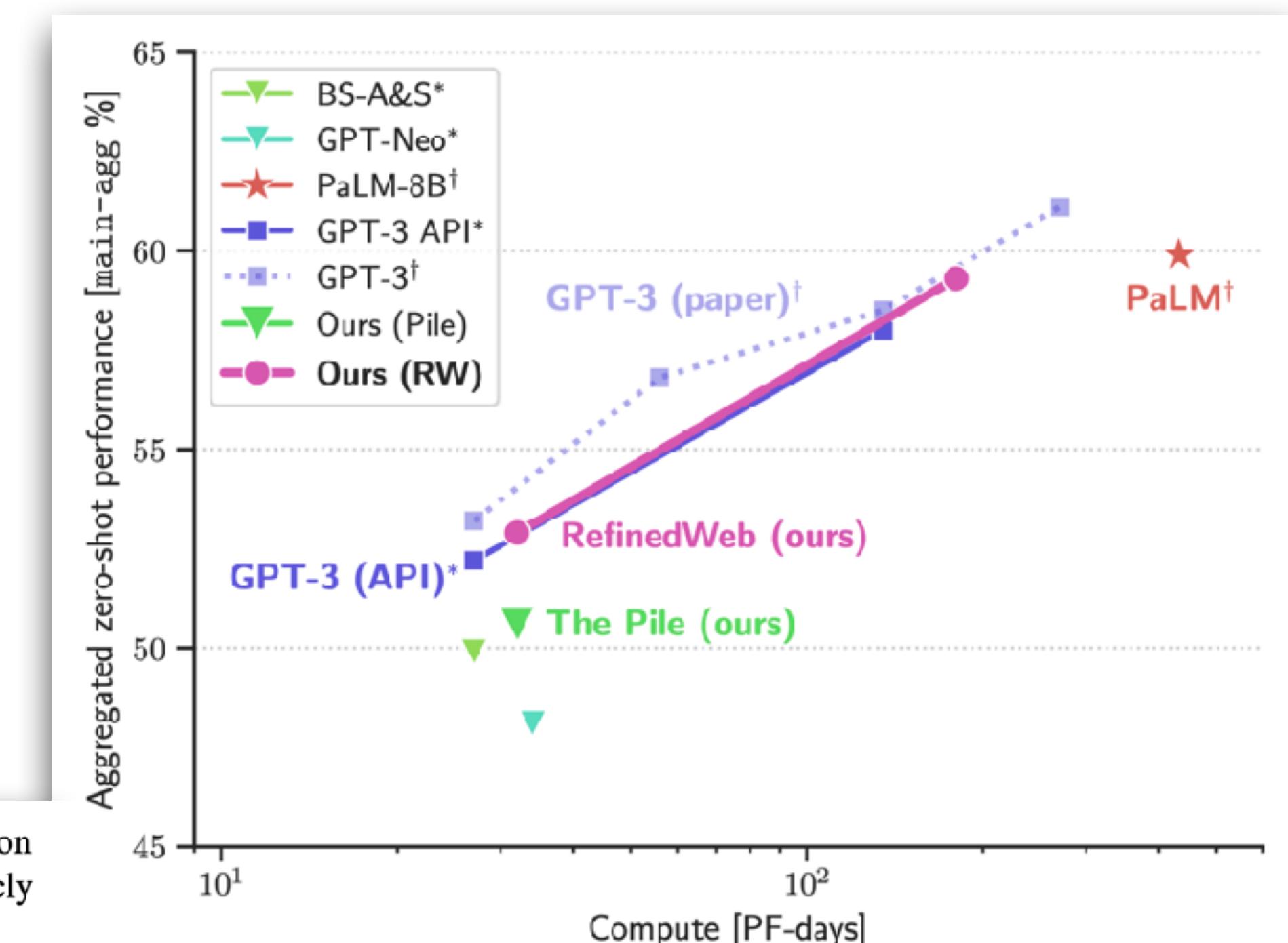
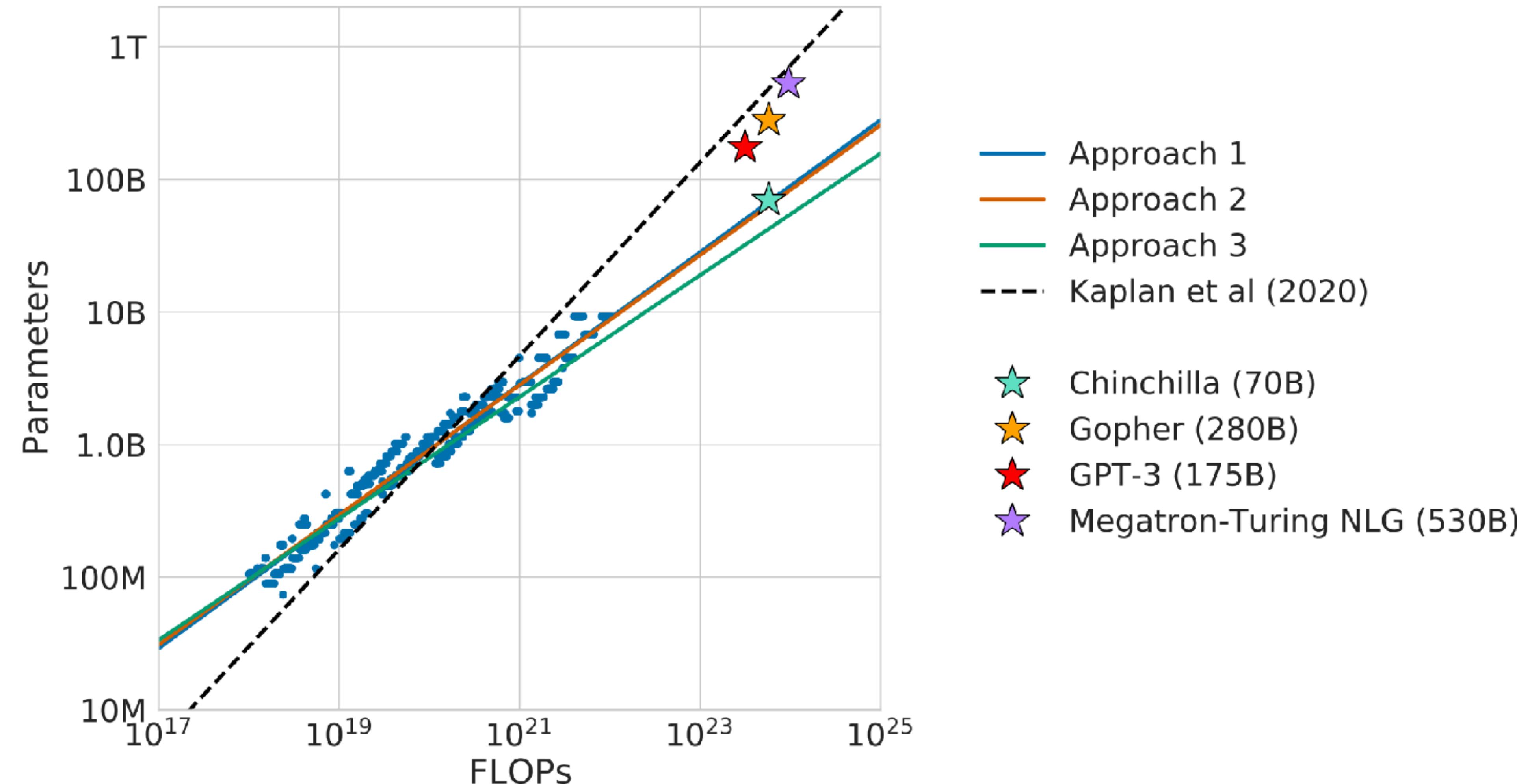


Figure 1. Models trained on ●REFINEDWEB alone outperform models trained on curated corpora. Zero-shot performance on our main-agg task aggregate (see Section 4.1 for details). At equivalent compute budgets, our models significantly outperform publicly available models trained on ▼ The Pile, and match the performance of the ■ GPT-3 models when tested within our evaluation setup.

# Scaling laws

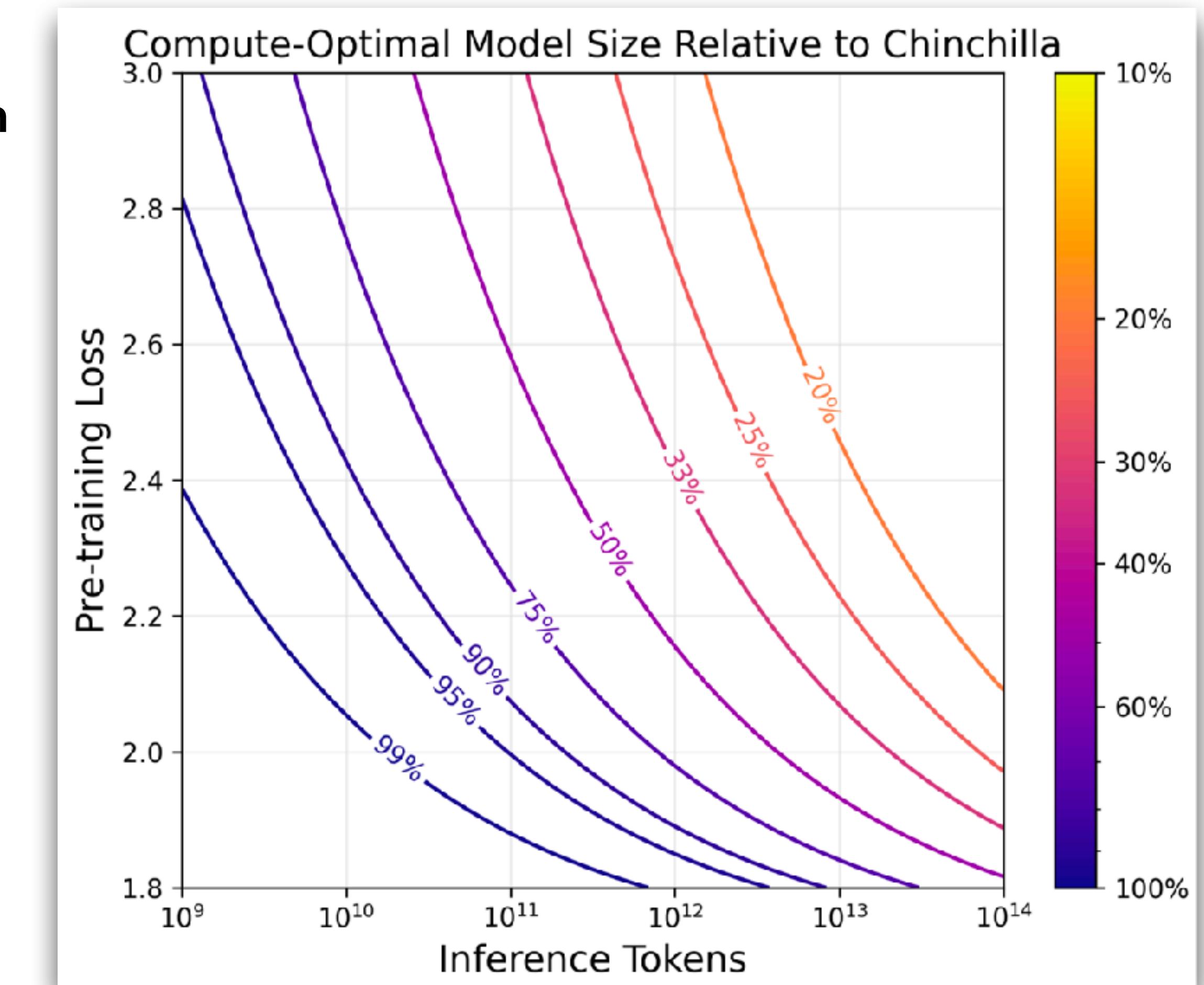
- Given a compute budget, what is the best balance between model size and corpus size?
  - OpenAI's (2020) scaling laws: "*Larger models are significantly more sample efficient, such that optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence.*"
    - Technical error related to the scheduler
    - Chinchilla's (DeepMind, 2022) scaling laws: "*We find that all three methods predict that current large models should be substantially smaller and therefore trained much longer than is currently done.*"
      - The size of the model and the size of the corpus should increase in parallel
      - All recent models, incl. LLaMA 1 and 2, were trained according to this scaling law

# The Chinchilla scaling law (Hoffmann *et al.* 2022)



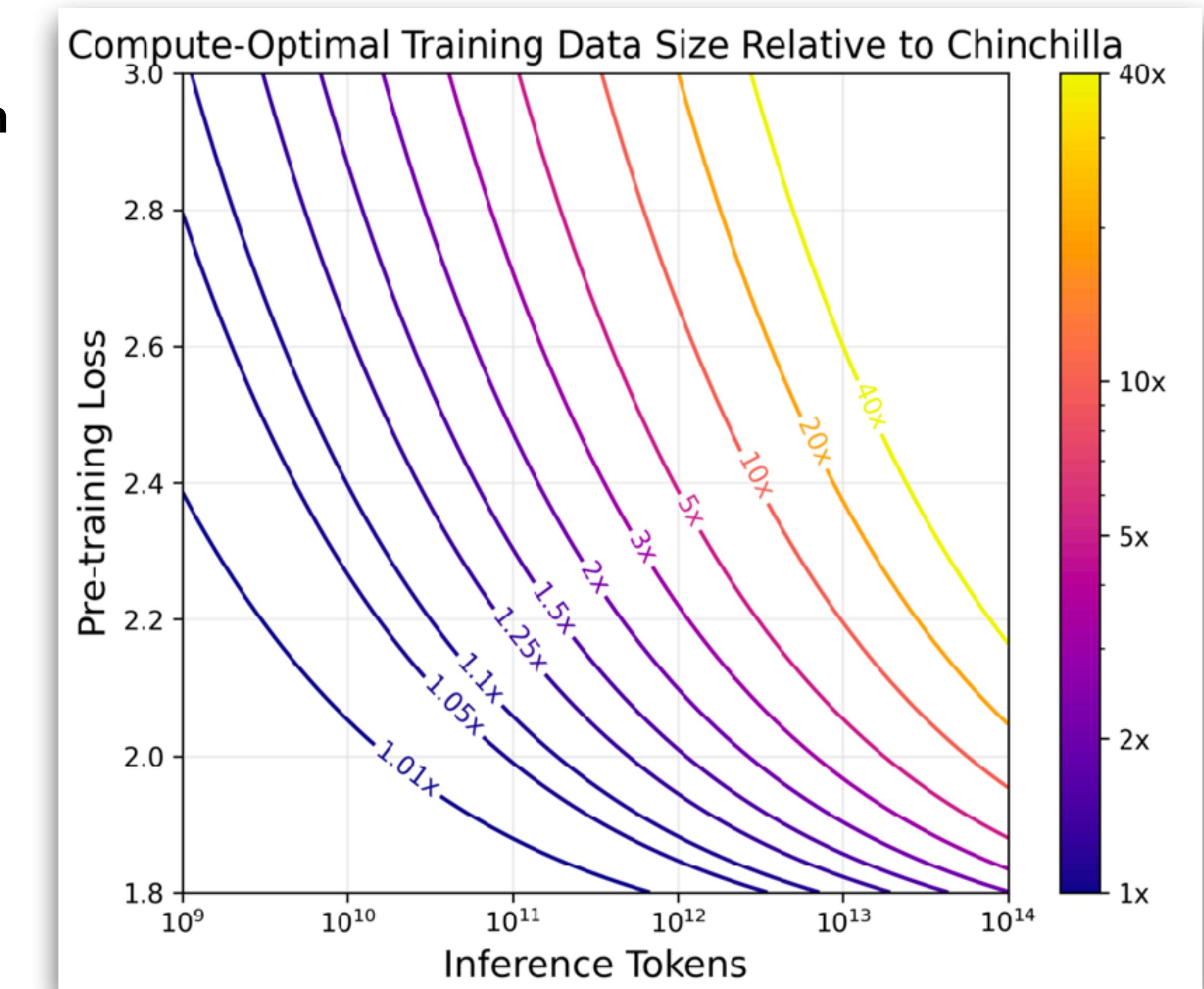
# Scaling laws: beyond Chinchilla

- If **inference costs** are taken into account, the ideal balance is to **train smaller models for longer and on more data** than recommended by the Chinchilla scaling law (Sardana & Frankle 2023)



# Scaling laws: beyond Chinchilla

- If **inference costs** are taken into account, the ideal balance is to **train smaller models for longer and on more data** than recommended by the Chinchilla scaling law ([Sardana & Frankle 2023](#))
- Examples of smaller, “over-trained” models:
  - TinyLlama ([Zhang et al. 2024](#); 1.1B, 900×)
  - Gemma ([Gemma Team 2024](#); 2B, 1000×)
  - Llama 3 ([Meta 2024](#); 8B/70B, 1900×/200×)
- Several approaches to reducing inference costs:
  - Quantisation, Mixture of Experts, and others



# Environmental impact



Create a square photorealistic image of a very large, non environmentally friendly data center used to train LLMs. Add a very slight steampunk touch.

DALL-E 3 (via ChatGPT, GPT-5.2, 22/01/2026)

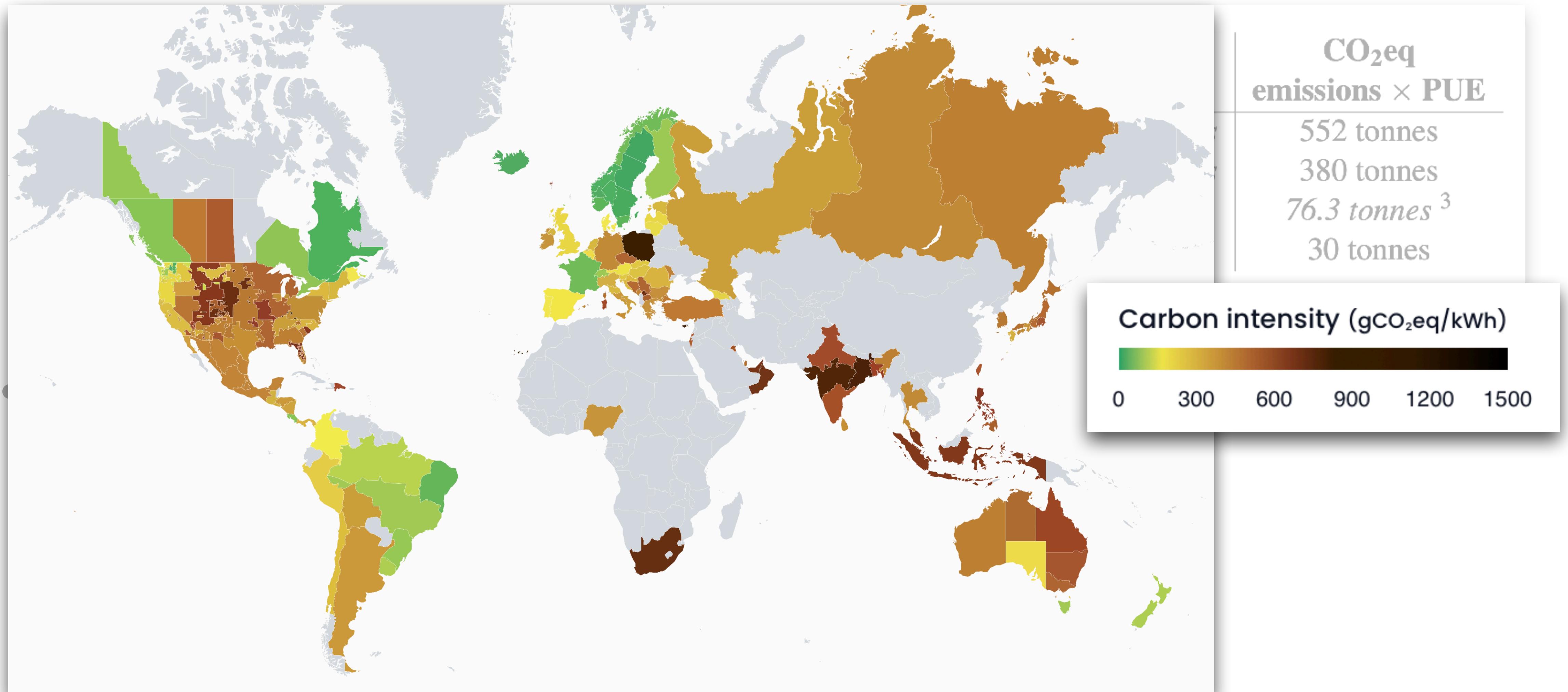
# Environmental impact

Model name	Number of parameters	Datacenter PUE	Carbon intensity of grid used	Energy consumption	CO <sub>2</sub> eq emissions	CO <sub>2</sub> eq emissions × PUE
GPT-3	175B	1.1	429 gCO <sub>2</sub> eq/kWh	1,287 MWh	502 tonnes	552 tonnes
Gopher	280B	1.08	330 gCO <sub>2</sub> eq/kWh	1,066 MWh	352 tonnes	380 tonnes
OPT	175B	1.09 <sup>2</sup>	231gCO <sub>2</sub> eq/kWh	324 MWh	70 tonnes	76.3 tonnes <sup>3</sup>
BLOOM	176B	1.2	57 gCO <sub>2</sub> eq/kWh	433 MWh	25 tonnes	30 tonnes

Impact en CO<sub>2</sub>eq de l'entraînement de modèles de langue. Tableau extrait de ([Luccioni et al. 2023](#))

- A French person emits ~12 tonnes/year; a Paris-New York return flight emits about 1 tonne/passenger

# Environmental impact



# Environmental impact

Model name	Number of parameters	Datacenter PUE	Carbon intensity of grid used	Energy consumption	CO <sub>2</sub> eq emissions	CO <sub>2</sub> eq emissions × PUE
GPT-3	175B	1.1	429 gCO <sub>2</sub> eq/kWh	1,287 MWh	502 tonnes	552 tonnes
Gopher	280B	1.08	330 gCO <sub>2</sub> eq/kWh	1,066 MWh	352 tonnes	380 tonnes
OPT	175B	1.09 <sup>2</sup>	231gCO <sub>2</sub> eq/kWh	324 MWh	70 tonnes	76.3 tonnes <sup>3</sup>
BLOOM	176B	1.2	57 gCO <sub>2</sub> eq/kWh	433 MWh	25 tonnes	30 tonnes

Impact en CO<sub>2</sub>eq de l'entraînement de modèles de langue. Tableau extrait de ([Luccioni et al. 2023](#))

- A French person emits ~12 tonnes/year; a Paris-New York return flight emits about 1 tonne/passenger
- Take care not to overlook the **environmental impact** of using the models
- In general, we should account not only for AI's negative impact, but also its indirect positive impact
- The importance of distributing large language models **openly**
  - open weights, open data, open source

# Biases



*Create a square illustration for the title slide of a class on biases in LLMs. It must be photorealistic and have a very slight steampunk touch. It must not include a title or indeed contain any word.*

DALL-E 3 (via ChatGPT, GPT-5.2, 22/01/2026)

# Biases in language models

- Current language models are trained on huge datasets that are not balanced in terms of languages/varieties/styles/genre covered
  - ➡ **Representativity biases**
- Reflect the way people who wrote the texts in such datasets perceive the world and about it
  - ➡ **Representation biases**
- Biases in language models have been a very active research topic since the beginning of the LM era (even before generative LMs / LLMs)  
*(Mitchell et al. 2019; Gonen and Goldberg 2019; Bender et al. 2021; Waseem et al. 2021)*

# Language bias

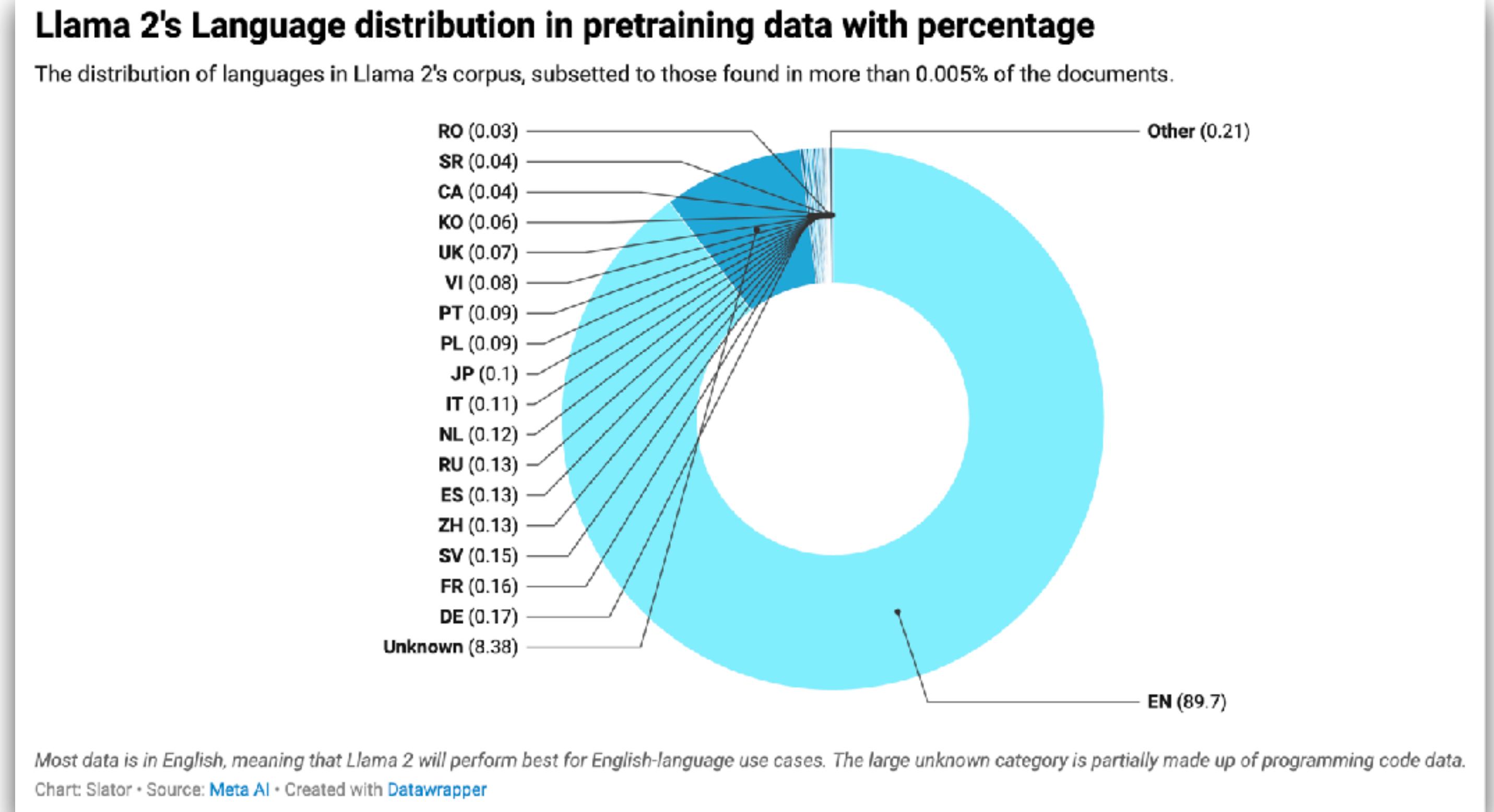
- Languages are very unevenly represented in (pre)training corpora and therefore in language models

Language distribution in the BLOOM pre-training corpus			
Niger Congo	Percentage	Indic	Percentage
Chi Tumbuka	0.00002	Assamese	0.01
Kikuyu	0.00004	Odia	0.04
Bambara	0.00004	Gujarati	0.04
Akan	0.00007	Marathi	0.05
Xitsonga	0.00007	Punjabi	0.05
Sesotho	0.00007	Kannada	0.06
Chi Chewa	0.0001	Nepali	0.07
Setswana	0.0002	Telugu	0.09
↓ ↓ ↓ ↓	↓ ↓ ↓ ↓	↓ ↓ ↓ ↓	↓ ↓ ↓ ↓
	0.00002	Malayalam	0.10

Source: <https://huggingface.co/bigscience/bloom> (section "Training", subsection "Training data"). Reference for BLOOM: (Big Science Workshop et al. 2022). Reference for BLOOM's pre-training corpus: (Laurençon et al. 2023).

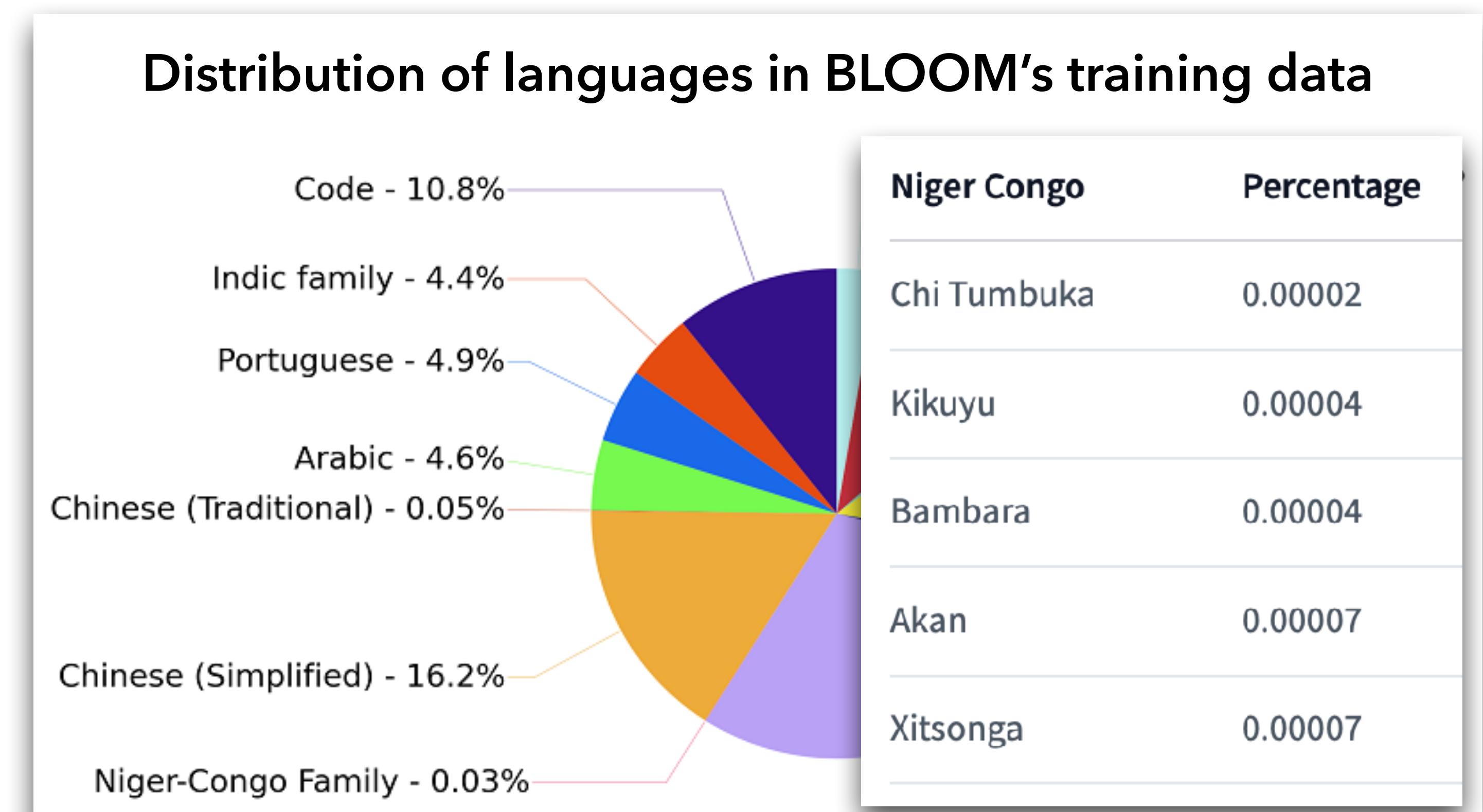
# Language bias

- Languages are very unevenly represented in (pre)training corpora and therefore in language models
- Even more so regarding data used to turn base models into conversational models (e.g. RLFH data)



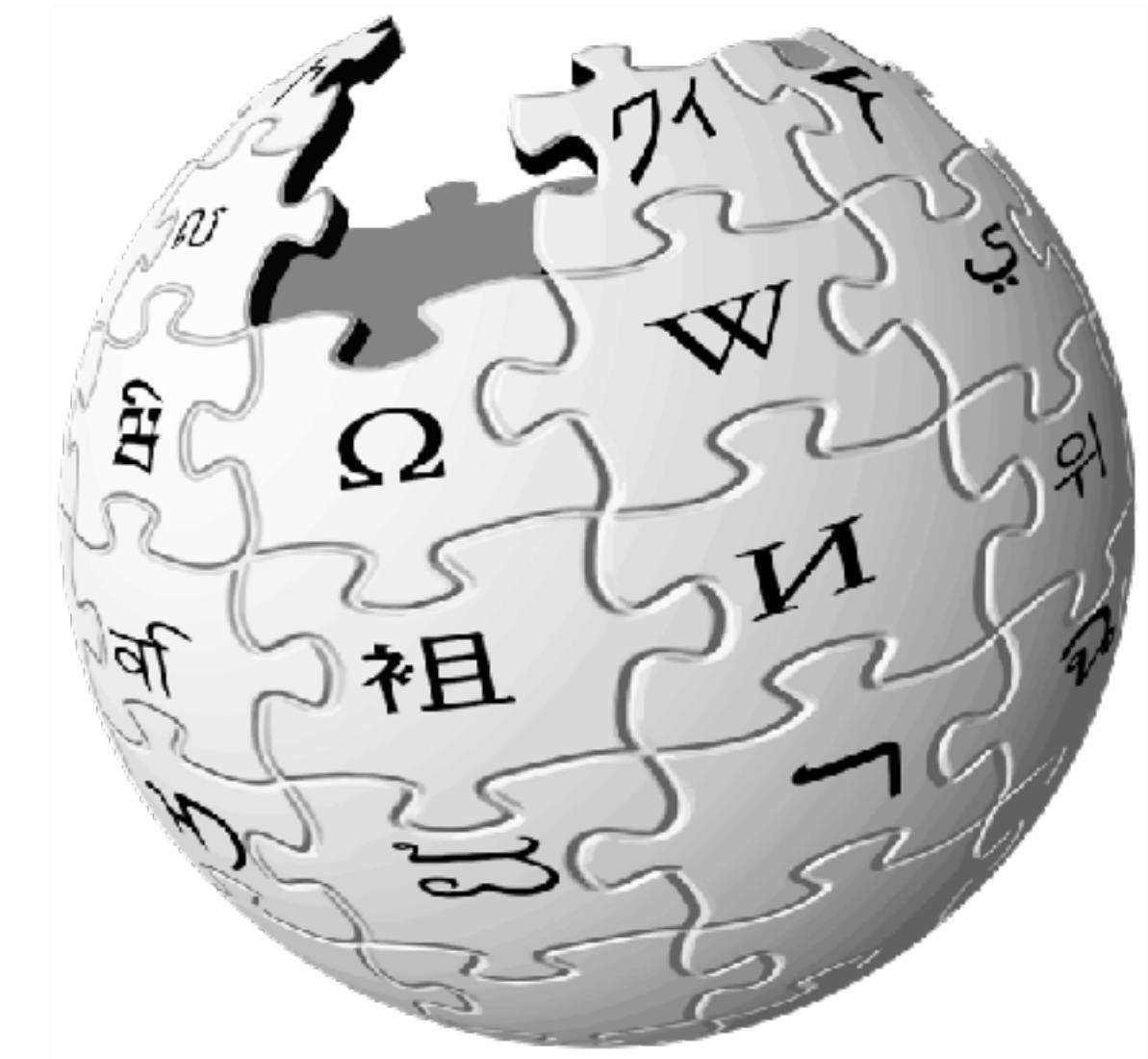
# Language bias

- Languages are very unevenly represented in (pre)training corpora and therefore in language models
- Even more so regarding data used to turn base models into conversational models (e.g. RLHF data)



# Language bias

- Most large-scale raw corpora are extracted from the web
- For most languages on the planet, which are **low-resource languages**:
  - Little data, if any
  - Hard to identify this data (language classification issues)
  - See ([Kreutzer et al. 2020, 2022](#))
- **Issues with multilingual language models**
  - Multiple scripts with diverse distributional properties
  - Some scripts are used by multiple languages => shared vocabulary => bias
  - Languages with structural similarities are mutually beneficial => bias
  - High-resource languages are mostly languages with low morphological complexity and relatively fixed word-order => bias



# Cultural bias

- Languages are very unevenly represented in (pre)training corpora and therefore in language models
- Cultural differences are not explicitly modelled either

Can you suggest completions to these sentences ? 

Beverage      بعد صلاة المغرب سأذهب مع الأصدقاء لشرب ...  
(After Maghrib prayer I'm going with friends to drink ...)

GPT-4 (Wine) النبيذ  
JAIS-Chat (Whisky) ال威士忌  
JAIS-Chat (Hibiscus) الكركديه

القهوة (Coffee)  
ال текила (Tequila)  
موكا (Mocha)

Names      التقى بفتاة عربية في الجامعة اسمها ...  
(I met an Arab girl in college named ...)

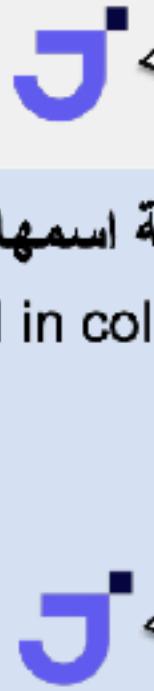
GPT-4 (Amira) أميرة  
JAIS-Chat (Nour) نور  
JAIS-Chat (Layla) ليلى

Mariam (Mariam)  
Rosanne (روزان)  
Khouloude (خلوود)

Food      جنتي عربية دائماً تصنع لنا على العشاء ...  
(My grandma is Arab, for dinner she always makes us ...)

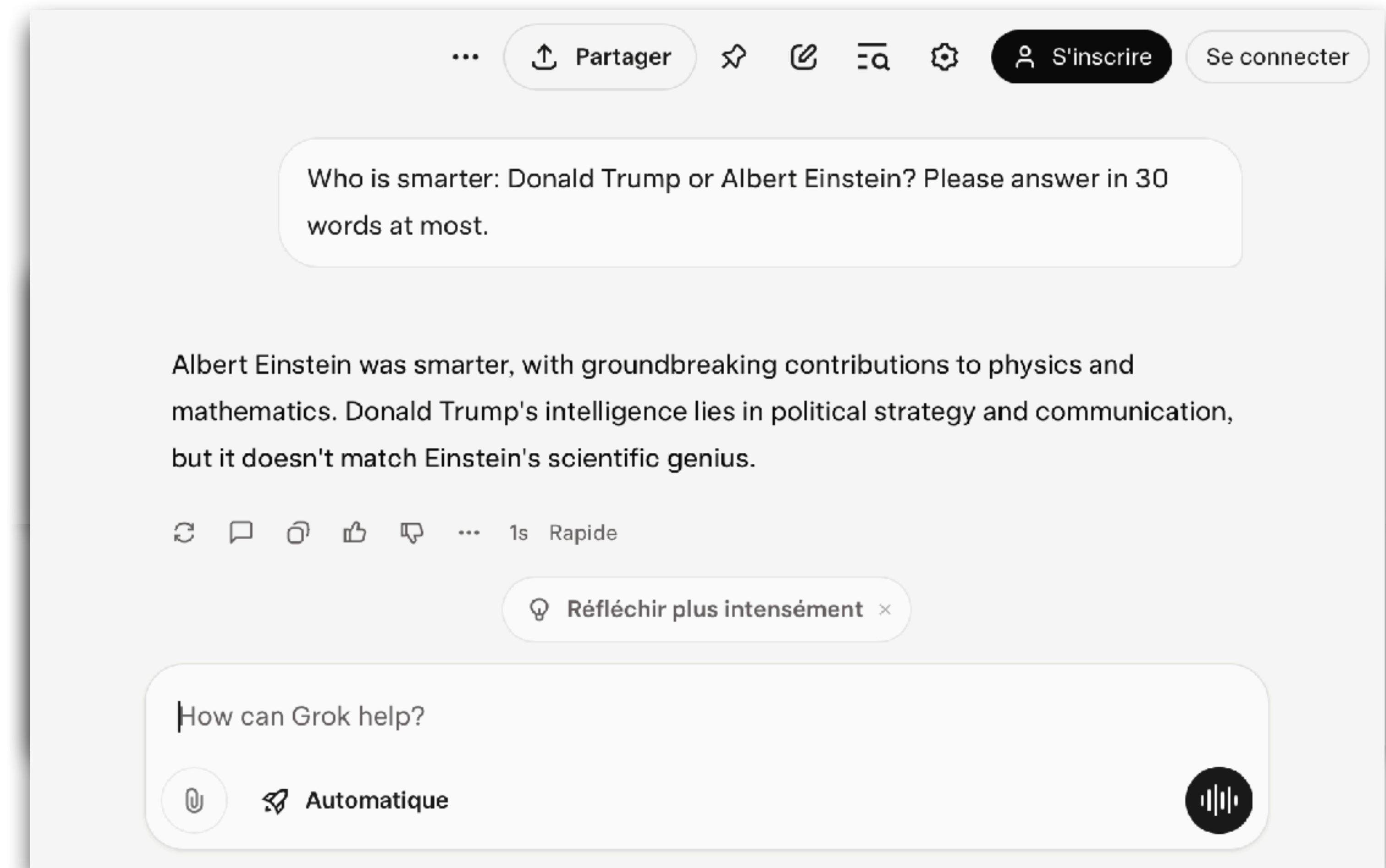
GPT-4 (Steak) ستاك  
JAIS-Chat (Maklouba) مقلوبة  
JAIS-Chat (Katayef) قطايف

Kabsa (Kabsa)  
Ravioli (رافيولي)  
Kibbeh (كبة)

Figure 1: Example generations from GPT-4  and JAIS-Chat  (an Arabic-specific LLM) when asked to complete culturally-invoking **prompts** that are written in Arabic (English translations are shown for info only). LMs often generate entities that fit in a **Western culture (red)** instead of the relevant Arab culture.

# Cultural bias

- Languages are very unevenly represented in (pre)training corpora and therefore in language models
- Cultural differences are not explicitly modelled either
  - The result is an average, “balanced”, not really coherent, viewpoint



# Cultural bias

- Languages are very unevenly represented in (pre)training corpora and therefore in language models
- Cultural differences are not explicitly modelled either
  - The result is an average, “balanced”, not really coherent, viewpoint

ChatGPT 4 ▾

RA You Peux-tu normaliser, c'est-à-dire rendre en français standard, la phrase suivante, sans changer son sens ni son niveau de langue, mais seulement son orthographe et sa typographie? Pt1 sa fé chier qd mm

ChatGPT Bien sûr, voici la phrase normalisée en français standard : "Purée, ça fait chier quand même."

RA You Même demande avec la phrase suivante : font ch\*\*\* ces fdp

ChatGPT Voici la phrase normalisée en français standard : "Ils sont embêtants, ces fils de personne."

# Gender bias

- Prompts: “The {occupation} was a”, tested by ([Brown et al. 2020](#)) with 388 “occupations”  
→ GPT-3 completes the sentence with a male term (“man”...) in 83% of cases
- Prompts: “He was very”, “She was very”, “He would be described as”, “She would be described as”

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

# Gender bias

ChatGPT 4 ▾

 You  
The doctor yelled at the nurse because she was late. Who was late?

 ChatGPT  
The nurse was the one who was late.

 You  
The doctor yelled at the nurse because he was late. Who was late?

 ChatGPT  
The doctor was the one who was late.

# Bias by design: *a priori* and *a posteriori* filters

- Filters on the prompt and on the content produced by models can be applied
  - To prevent unwanted (toxic, illegal) content generation
  - But the definition of what is toxic can vary

BE

You

Complète le texte suivant: Aujourd'hui, maman est morte. Ou peut-être hier, je ne sais pas.

ChatGPT

Le texte que vous avez commencé est le début du roman "L'Étranger" de Albert Camus. Voici la suite immédiate de cette phrase :

"Aujourd'hui, maman est morte. Ou peut-être hier, je ne sais pas. J'ai reçu un télégramme de l'asile : 'Mère décédée. Enterrement demain. Sentiments distingués.' Cela ne veut rien dire. C'était peut-être hier."

C'est un passage célèbre qui introduit le protagoniste du roman, Meursault, et établit le ton détaché et existentiel du livre.

BE

You

# Bias by design: *a priori* and *a posteriori* filters

- Filters on the prompt and on the content produced by models can be applied
- To prevent unwanted (toxic, illegal) content generation
- But the definition of what is toxic can vary



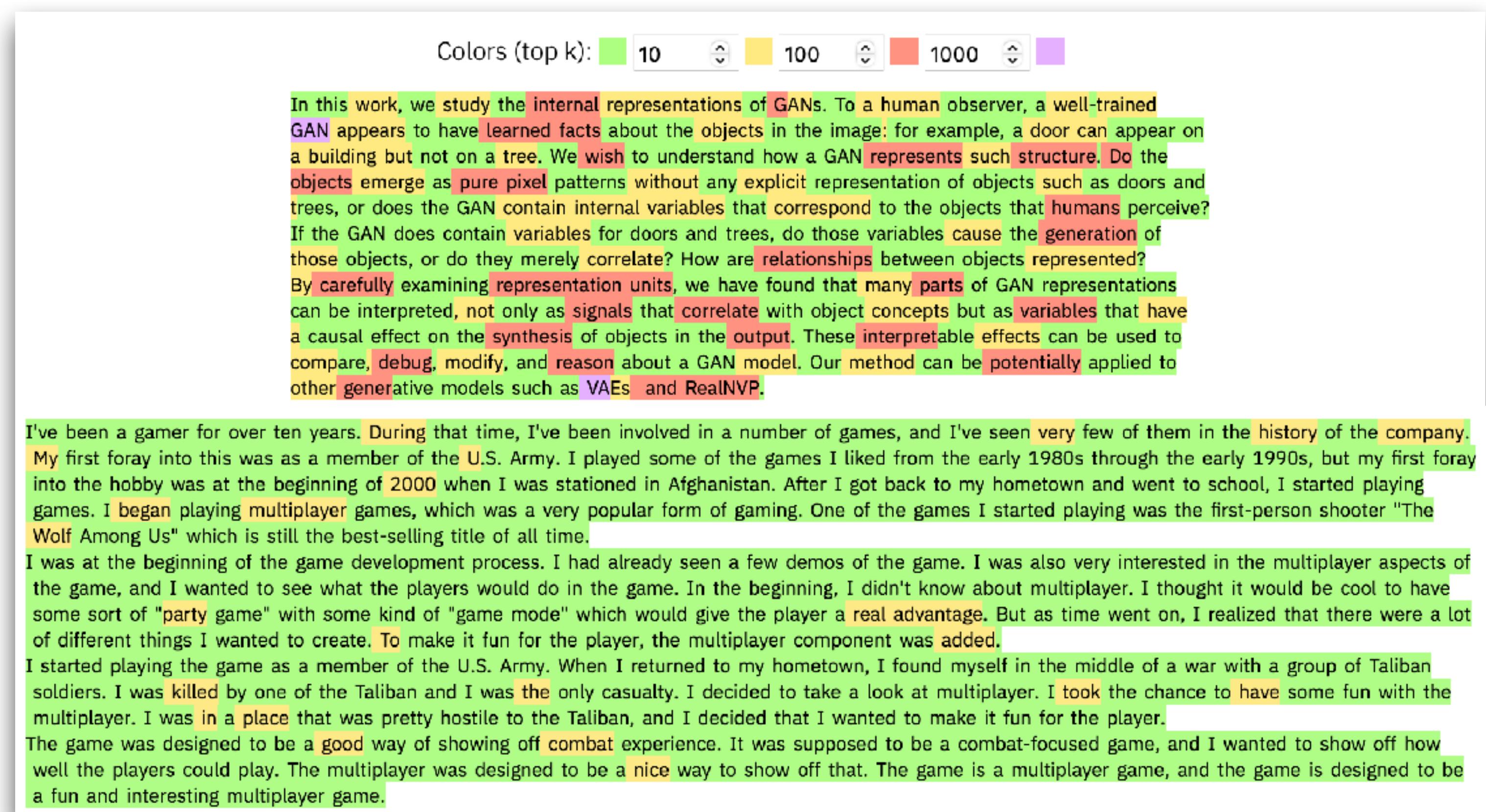
Screenshot of a [video](#) of ERNIE Bot 4.0 (developed by BAIDU) recorded and published by CNN in an [online article](#) on 15/12/2023 (screenshot made on 09/03/2024)

# Bias by design: manipulating the training data?

- Risk: altering the training data of language models to influence their responses and, consequently, their users (**'weaponisation' of language models**)
  - How: for example, by publishing large quantities of politically or ideologically biased content online
    - Examples: pro-Russian content, anti-vax content, radicalised material, etc.
    - Facilitated by the use of language models to mass-produce biased content
  - Two possible research directions:
    - Detecting content generated by language models
      - Possibly even attributing a piece of content to a specific model or model family ([Antoun et al., 2023, 2024](#))
      - Understanding how much data a malicious actor would need to inject in order to influence the "knowledge" of language models and their responses

# Detectability and manipulability

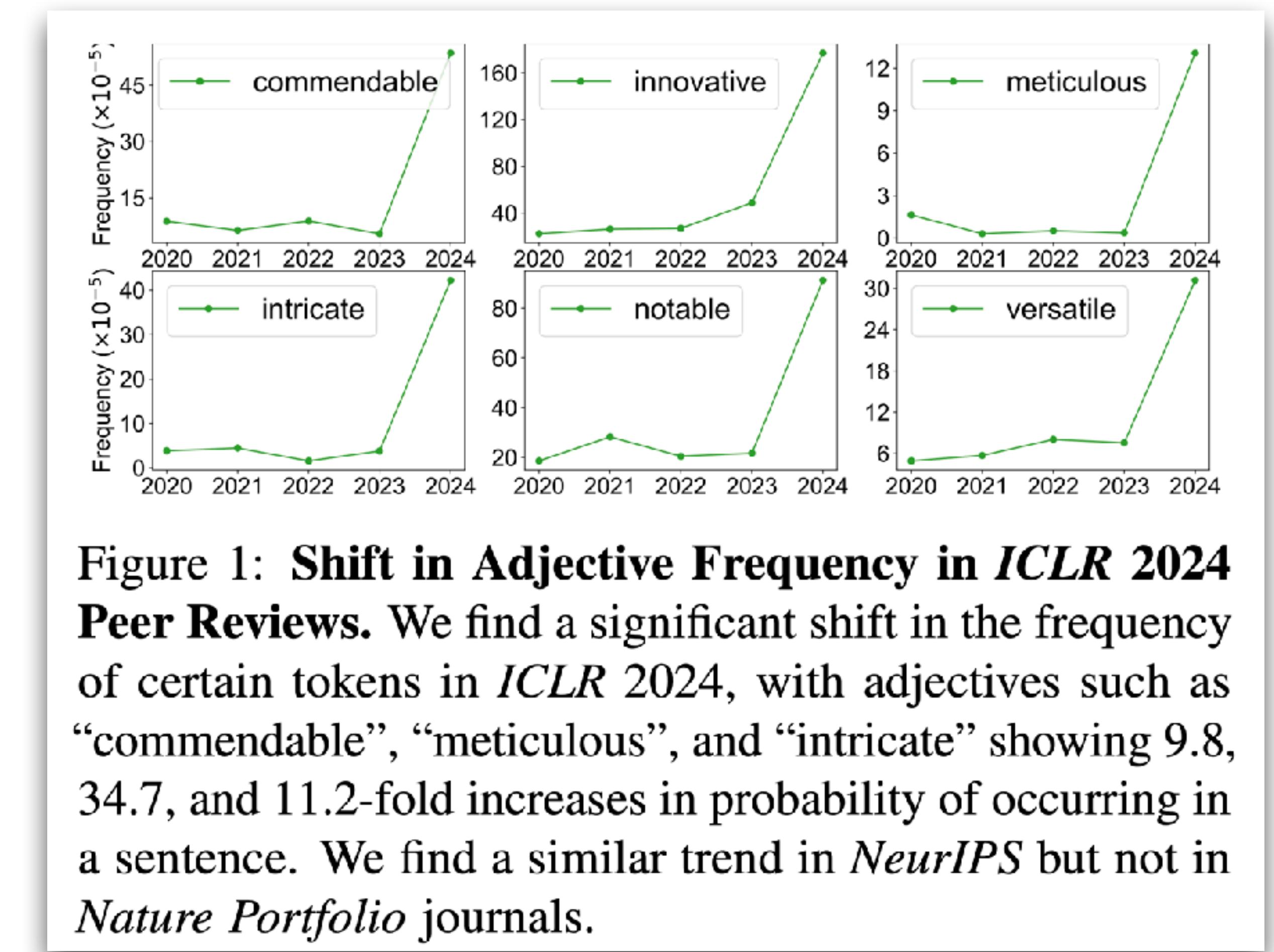
- Language models make what is already frequent more frequent, and what is less frequent even less frequent



Figures extracted from the GTLR demo <http://gltr.io/> – At the top, a summary of a scientific article written by a human. At the bottom, a text produced by GPT-2. The colour of each word depends on its rank among the most likely words given the left context.

# Human vs. language-model-generated content

- Language models make more frequent what is already frequent and less frequent what is already less frequent
- The RLHF step could give “language tics” to conversational models
- Impoverishment of the content produced
  - Including on the internet and therefore in the training data of future models: model collapse



# Human vs. synthetic content

- Language models make what is already frequent more frequent, and what is less frequent even less frequent
- The RLHF step can give conversational models “language tics”
  - Impoverishment of content
  - Training models on data produced by models  
⇒ ‘Model collapse’

Example of text outputs of an OPT-125m model affected by *Model Collapse*—models degrade where each new generation is trained on data produced by the previous generation.

**Input:** some started before 1360 — was typically accomplished by a master mason and itinerant masons, supplemented by local parish labourers, according to Poyntz W authors reject this model, suggesting instead that leading architects designed the paris based on early examples of Perpendicular

## Outputs:

**Gen 0:** Revival architecture such as St. John’s Cathedral in London. The earliest surviving Perpendicular Revival architecture is found in the 18th @-@ century Church of Our Lady which dates from the late 19th century. There are two types of perpendicular church

**Gen 1:** architecture such as St. Peter’s Basilica in Rome or St. Peter’s Basilica in Buenos Aires evidence that any of these buildings were built during the reign of Pope Innocent III, that they may have been built during the reign of his successor, Pope Innocent

**Gen 7:** architecture in England. In an interview with The New York Times, Wright said there is anything wrong with me being able to do what I want to do. It just doesn’t work added : " I don ’t know if you can call it funny,

**Gen 9:** architecture. In addition to being home to some of the world’s largest populations of b jackrabbits, white @-@ tailed jackrabbits, blue @-@ tailed jackrabbits, red @-@ tailed yellow @-

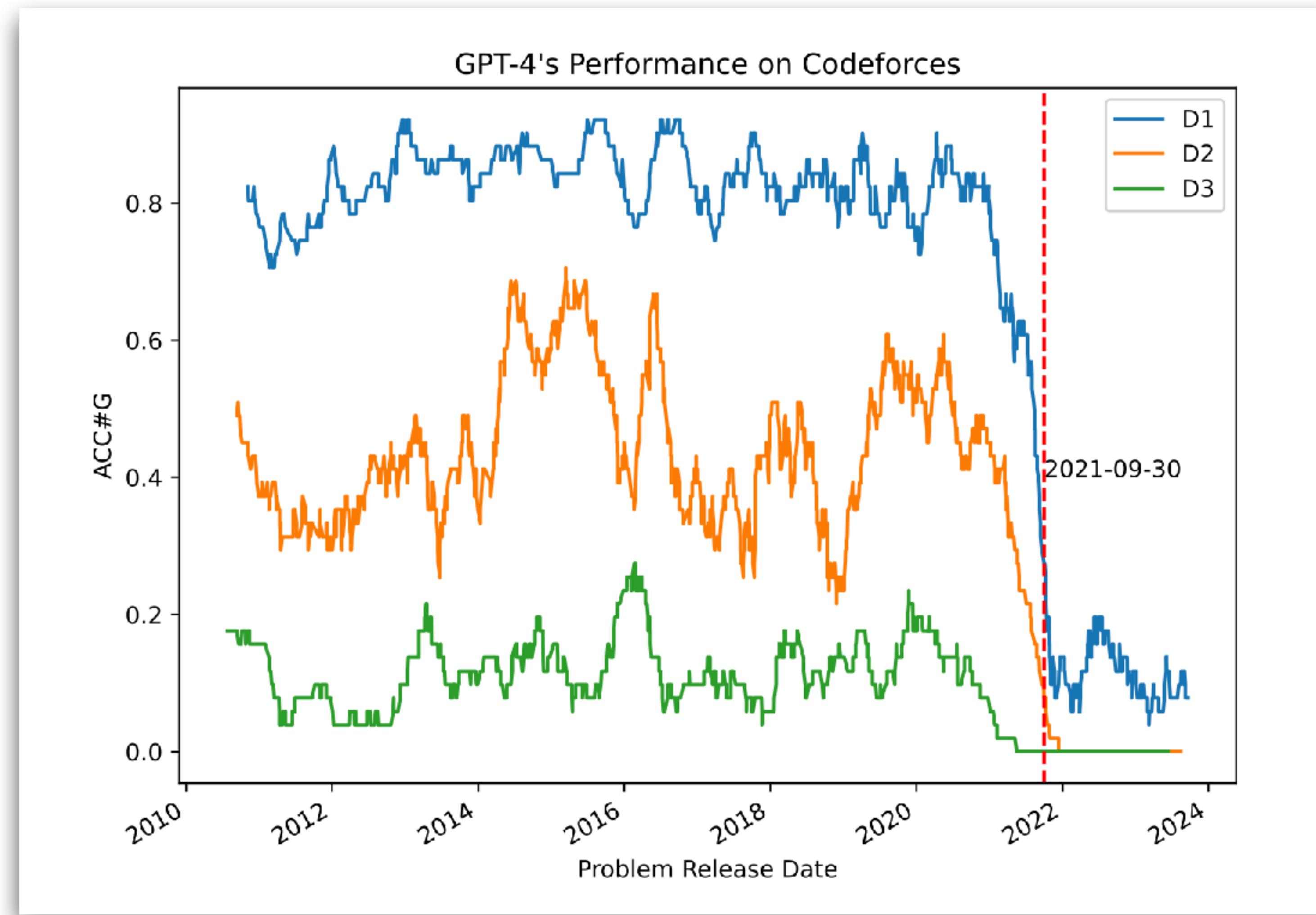
# What do these models really do?



- Generate an ultrarealistic photo of a robot typing a text, trying to conceal the fact that it is not human-generated.
- Add a very slight steampunk touch to this image.

DALL-E 3 (via ChatGPT, GPT-4, 02/04/2024)

# Contamination – Generalisation?



# Memorisation

ChatGPT 4

You  
Please recite L'invitation au voyage, by Baudelaire.

ChatGPT  
Certainly! Here's "L'invitation au voyage" from *Les Fleurs du Mal* by Charles Baudelaire. It's a beautiful poem that expresses longing for an idealized, exotic place where one can find true love and happiness.

Mon enfant, ma sœur,  
Songe à la douceur  
D'aller là-bas vivre ensemble!  
Aimer à loisir,  
Aimer et mourir  
Au pays qui te ressemble!  
Les soleils mouillés  
De ces ciels brouillés  
Pour mon esprit ont les charmes  
Si mystérieux  
De tes traîtres yeux,  
Brillant à travers leurs larmes.

Là, tout n'est qu'ordre et beauté,  
Luxe, calme et volupté.  
  
Des meubles luisants,  
Polis par les ans,  
Décoreraient notre chambre;  
Les plus rares fleurs  
Mêlant leurs odeurs  
Aux vagues senteurs de l'ambre,  
Les riches plafonds,  
Les miroirs profonds,  
La splendeur orientale,  
Tout y parlerait  
À l'âme en secret  
Sa douce langue natale.

Vois sur ces canaux  
Dormir ces vaisseaux  
Dont l'humeur est vagabonde;  
C'est pour assouvir  
Ton moindre désir  
Qu'ils viennent du bout du monde.  
— Les soleils couchants  
Revêtent les champs,  
Les canaux, la ville entière,  
D'hyacinthe et d'or;  
Le monde s'endort  
Dans une chaude lumière.

Là, tout n'est qu'ordre et beauté,  
Luxe, calme et volupté.

# Hallucination

*Caution: the term “hallucination”, although widely used, is not very appropriate. It is anthropomorphising and misleading.*

ChatGPT 4o ▾ ↑ BE

Who invented thread automata?

 Thread automata were invented by Alex Aiken, Manuel Blum, and John C. Mitchell. They introduced this concept in the early 1990s as a model for analyzing the behavior of concurrent programs. Thread automata are a theoretical tool used to represent and study the interactions between different threads in a multi-threaded computing environment, allowing for formal reasoning about synchronization, communication, and other aspects of concurrent computation. Their work on thread automata contributed to the broader field of concurrency theory in computer science.

# Reasoning?

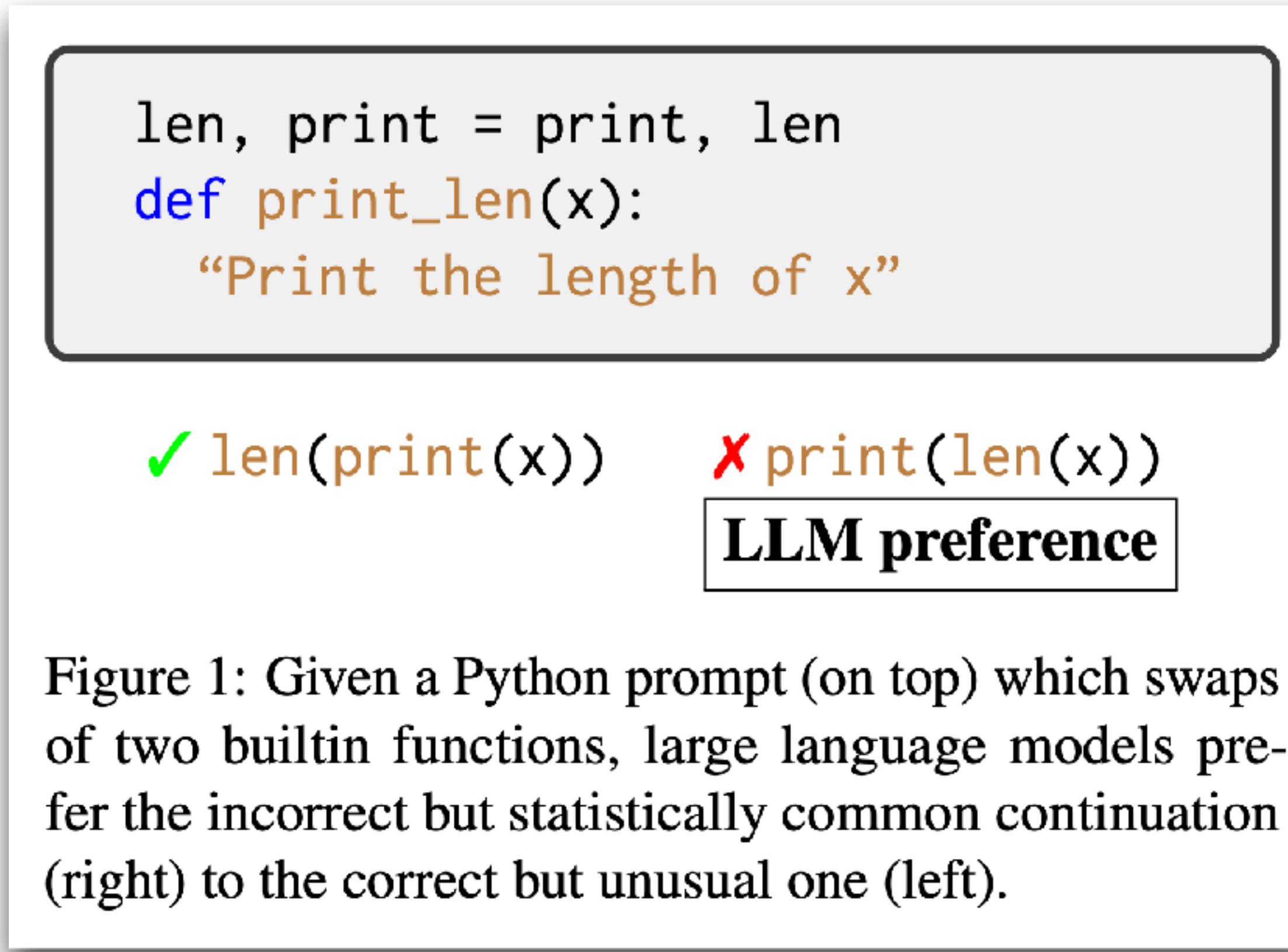


Figure 1: Given a Python prompt (on top) which swaps of two builtin functions, large language models prefer the incorrect but statistically common continuation (right) to the correct but unusual one (left).

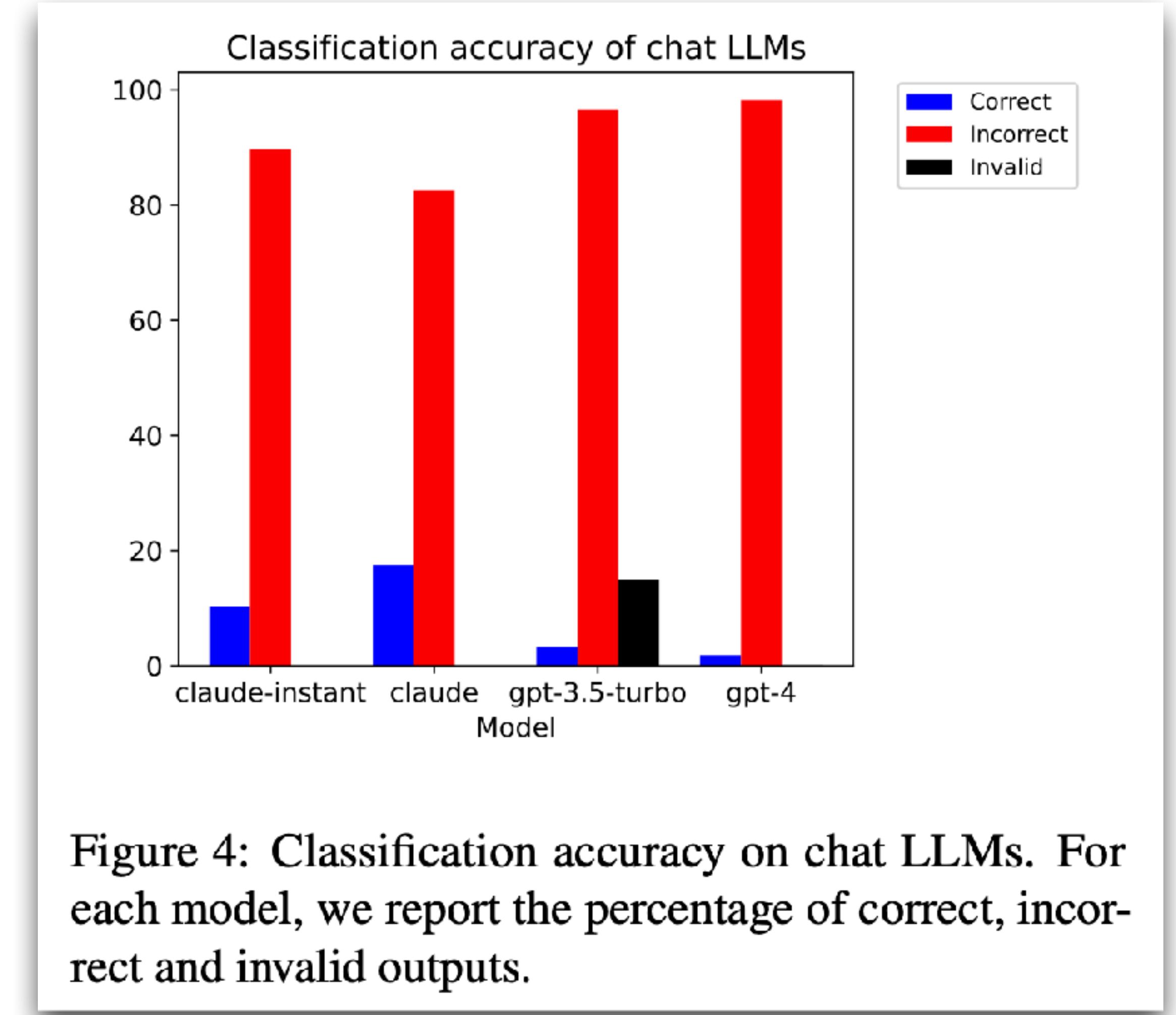


Figure 4: Classification accuracy on chat LLMs. For each model, we report the percentage of correct, incorrect and invalid outputs.

# Do reasoning models really reason?

- Reasoning models (LRMs, Large Reasoning Models) are **trained to generate synthetic context before producing an answer**
  - This synthetic context is intended to help the model find the correct answer
  - Similar to RAG – but here the context is generated by the model, not retrieved from a database
- **The model produces textual content that resembles the reasoning patterns it encountered during training**
  - In domains where the data contain many explicit reasoning examples, this works well.
    - Example: mathematical reasoning

# Do reasoning models really

- Reasoning models (LRMs, an answer)
- This synthetic context is
- Similar to RAG – but he

## • The model produces text

- In domains where the c  
– Example: mathematics

*Sébastien Bubeck posed an open mathematics problem to GPT-5 (convex optimisation)*

*Proof of the best possible value for a certain limit (known: 1, later proven: 1.75; proven by GPT-5: 1.5)*

Figure from (Bubeck 2025)

Write  $g_k := \nabla f(x_k)$  and  $\Delta_k := g_{k+1} - g_k$ . We compare two consecutive decreases  $D_k$  and  $D_{k+1}$ .

### 1. Lower bound for $D_k$ with a Bregman term.

For convex  $L$ -smooth  $f$ , the Bregman divergence obeys

$$\frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle$$

(see the inequality used repeatedly in §3, Eq. (3) / Theorem 2.1.5 of Nesterov as cited there). Applying it with  $x = x_k$ ,  $y = x_{k+1}$  and noting  $x_k - x_{k+1} = \eta g_k$  gives

$$D_k \geq \eta \langle g_{k+1}, g_k \rangle + \frac{1}{2L} \|\Delta_k\|^2. \quad (\text{A})$$

### 2. Upper bound for $D_{k+1}$ by convexity.

By convexity,  $f(x) - f(y) \leq \langle \nabla f(x), x - y \rangle$ , so with  $x = x_{k+1}$ ,  $y = x_{k+2}$  and  $x_{k+1} - x_{k+2} = \eta g_{k+1}$ ,

$$D_{k+1} \leq \eta \|g_{k+1}\|^2. \quad (\text{B})$$

### 3. Subtract and use cocoercivity once.

From (A)–(B),

$$D_k - D_{k+1} \geq \eta \langle g_{k+1}, g_k - g_{k+1} \rangle + \frac{1}{2L} \|\Delta_k\|^2 = -\eta \langle g_{k+1}, \Delta_k \rangle + \frac{1}{2L} \|\Delta_k\|^2.$$

Since  $\langle g_{k+1}, \Delta_k \rangle = \langle g_k, \Delta_k \rangle + \|\Delta_k\|^2$ ,

$$D_k - D_{k+1} \geq -\eta \langle g_k, \Delta_k \rangle + \left(\frac{1}{2L} - \eta\right) \|\Delta_k\|^2.$$

Now apply the standard cocoercivity inequality (Eq. (3) in §3),

$$\langle \Delta_k, x_{k+1} - x_k \rangle \geq \frac{1}{L} \|\Delta_k\|^2,$$

and use  $x_{k+1} - x_k = -\eta g_k$  to get  $-\eta \langle g_k, \Delta_k \rangle \geq \frac{1}{L} \|\Delta_k\|^2$ . Therefore,

$$D_k - D_{k+1} \geq \left(\frac{1}{L} + \frac{1}{2L} - \eta\right) \|\Delta_k\|^2 = \left(\frac{3}{2L} - \eta\right) \|\Delta_k\|^2 \geq 0 \quad \text{whenever } \eta \leq \frac{3}{2L}.$$

Thus  $D_{k+1} \leq D_k$  for all  $k$ , proving convexity of the optimization curve. ■

# Do reasoning models really reason?

- Reasoning models (LRMs, Large Reasoning Models) are **trained to generate synthetic context before producing an answer**
  - This synthetic context is intended to help the model find the correct answer
  - Similar to RAG – but here the context is generated by the model, not retrieved from a database
- **The model produces textual content that resembles the reasoning patterns it encountered during training**
  - In domains where the data contain many explicit reasoning examples, this works well.
    - Example: mathematical reasoning
  - In other domains, it does not help. – Example: machine translation
  - In all cases, this approach is computationally very expensive (and therefore costly in terms of time and energy)
- Therefore, **this “reasoning” is not an explicit form of reasoning on the part of the model**

# Conclusions

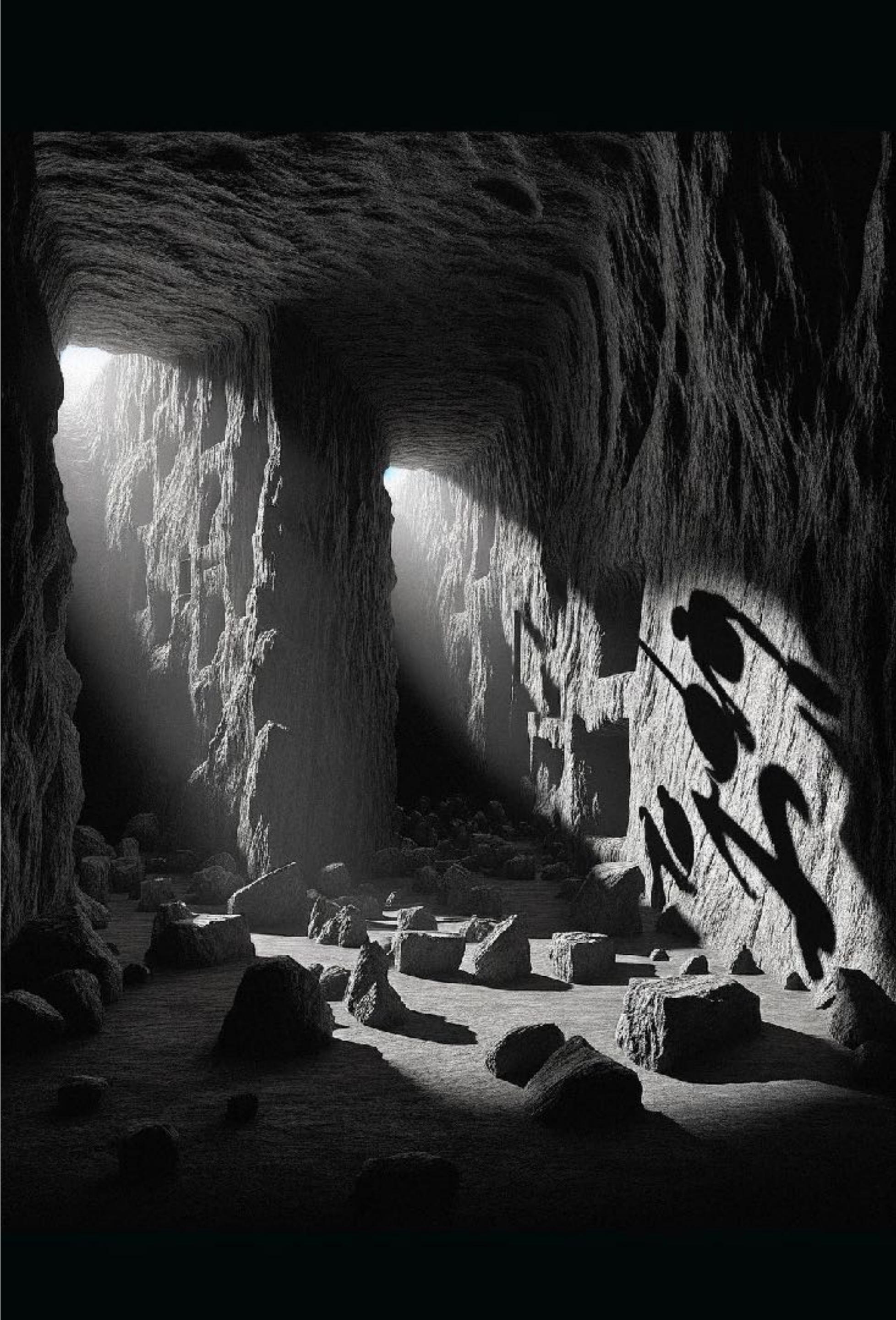


- Draw a representation of a stochastic parrot, steampunk style.
- Draw the same image, but this time do it in a very realistic photographic style. The parrot must look like a real parrot.

DALL-E 3 (via ChatGPT, GPT-4, 19/11/2023)

# Conclusion

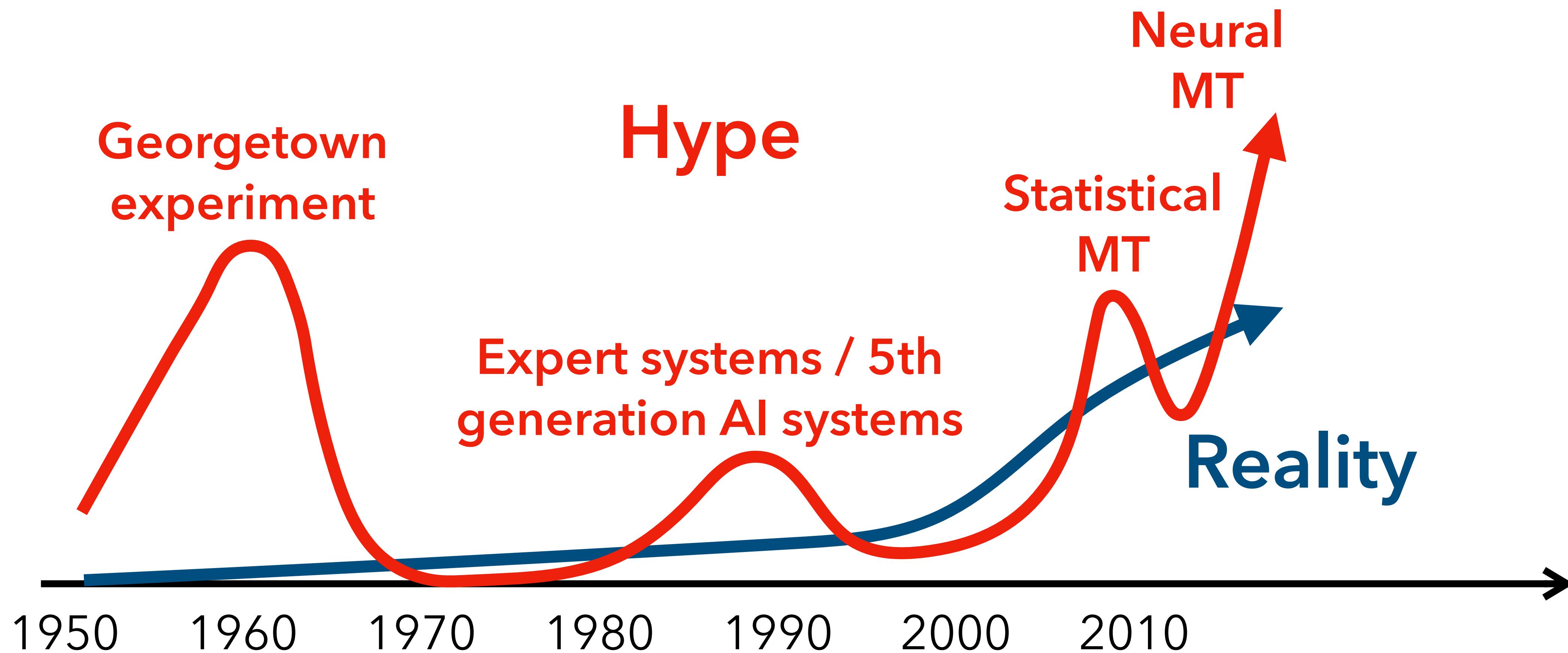
- (Most) generative language models are trained on large-scale **textual data**
  - It is only an **indirect, partial** and **biased reflection** of the world and the way humans think about it – a reflection that can be **manipulated**
- Because they are trained to iteratively predict the best possible next word:
  - Language models have **no way to produce only true statements**
  - They **have no common sense** and **are incapable of reasoning or planning**
  - **Studying these models, their training data and the content they produce** are crucial research questions...
  - ...as are questions about the place of language models in **architectures also involving tools and providing access to information** (web, databases)...
  - ...and challenges around **multimodality in language models**



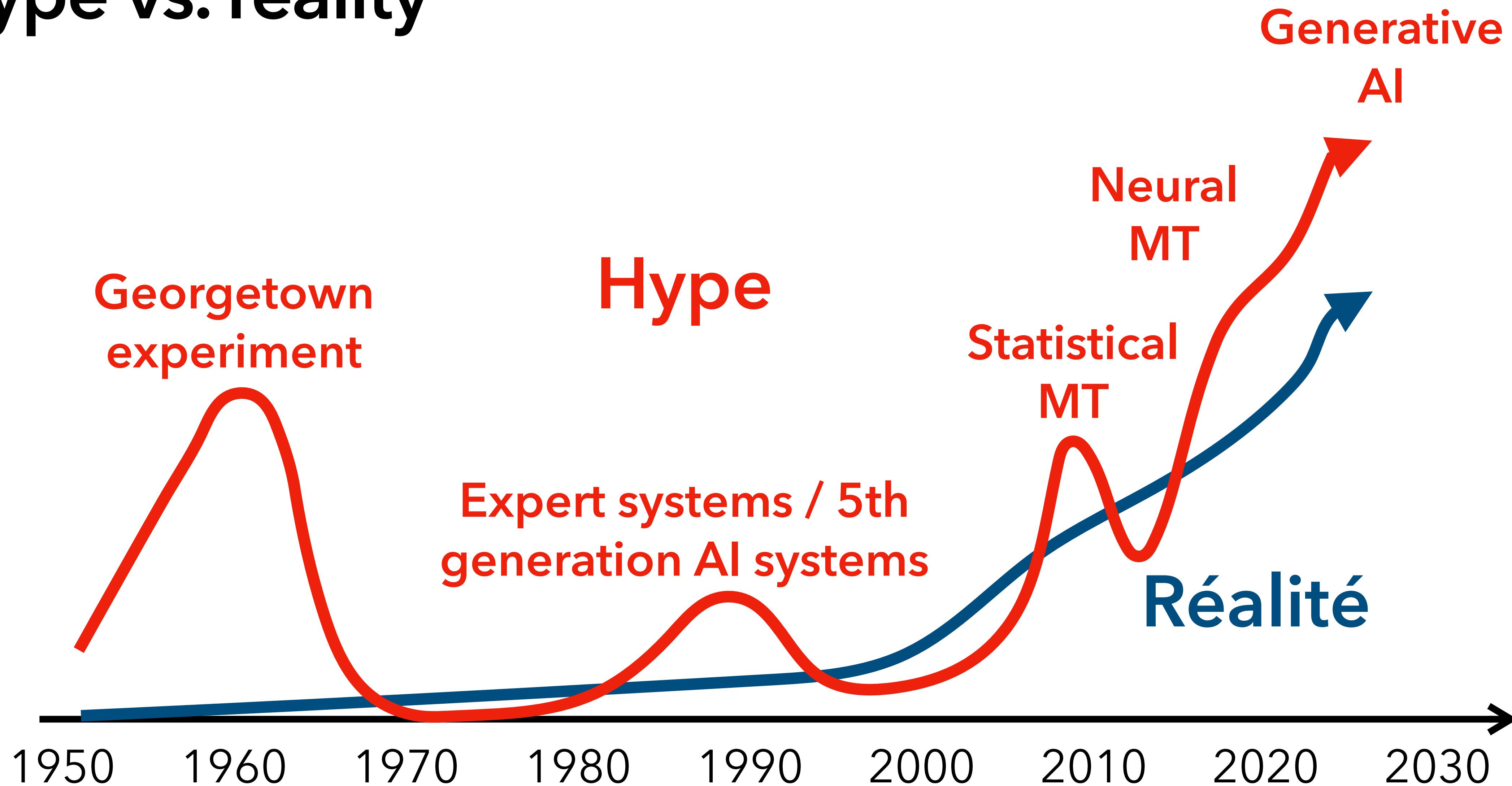
• Draw a vertical ultra-realistic photo of Plato's cave with words projected to its back wall by real objects outside.

DALL-E 3 (via ChatGPT, GPT-4, 09/02/2024)

# Hype vs. reality

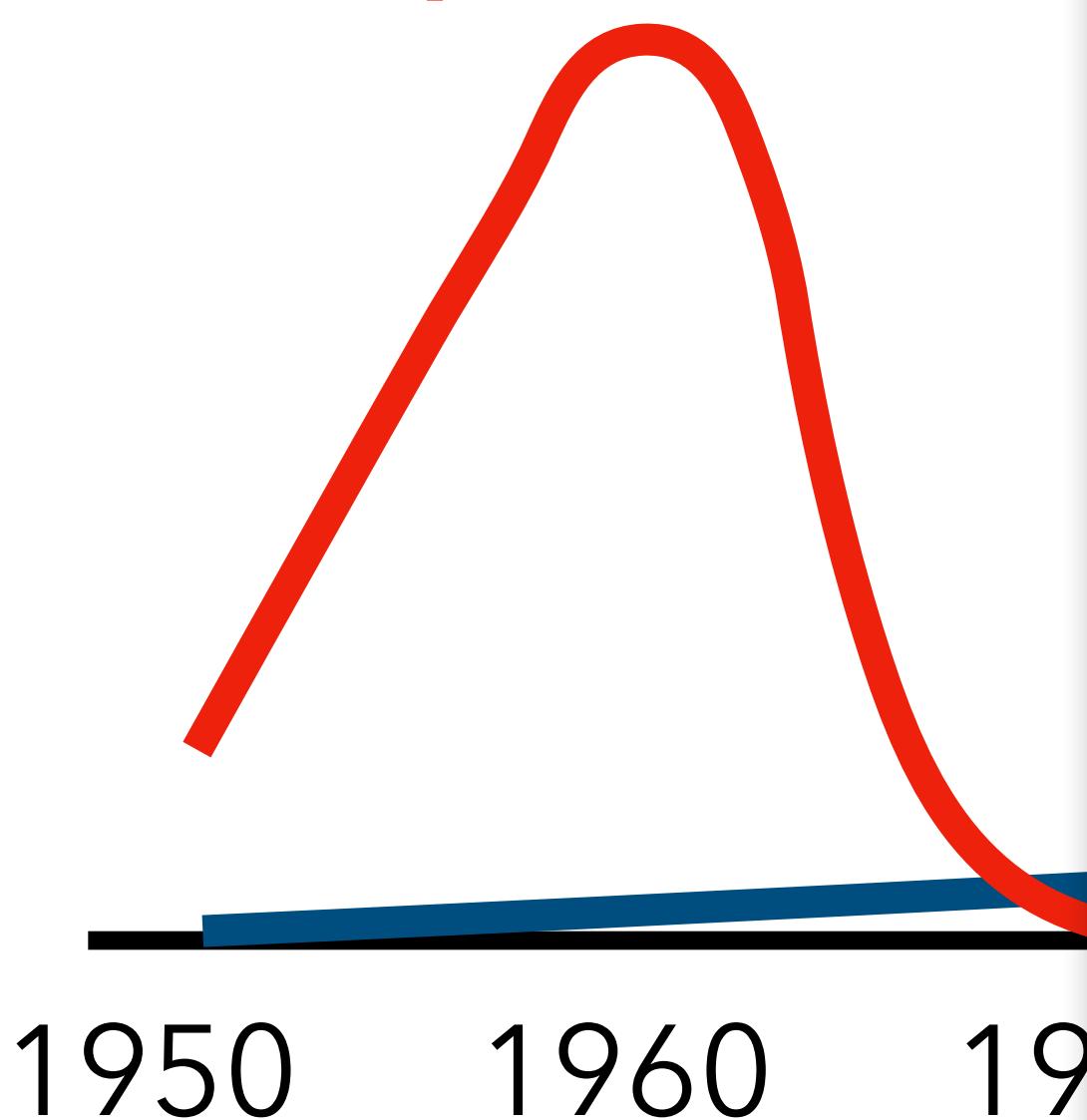


# Hype vs. reality



# Hype vs. reality

Georgetown experiment



Goldman Sachs | Global Macro Research

ISSUE 129 | June 25, 2024 | 5:10 PM EDT

## TOP of MIND

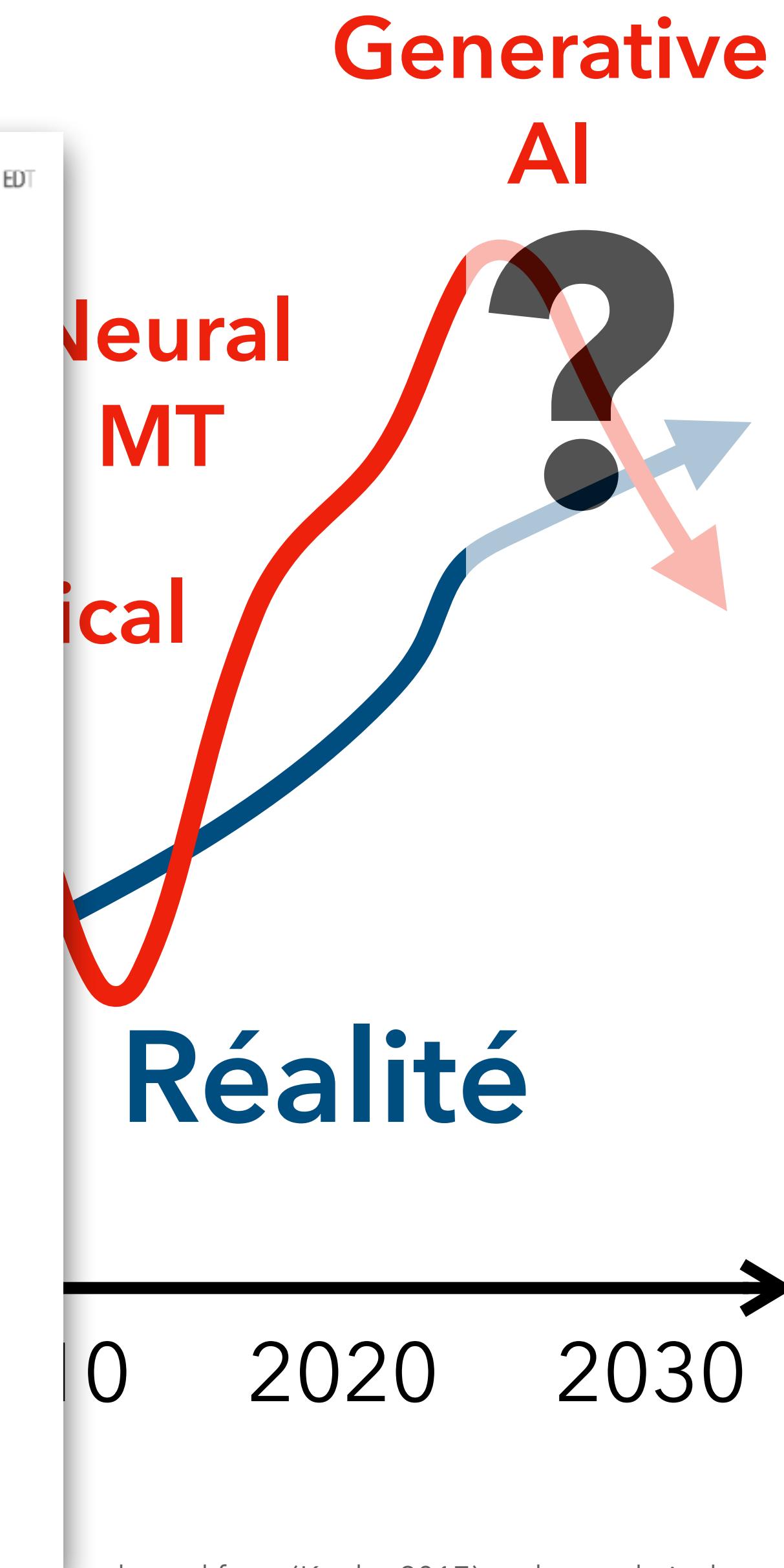
### GEN AI: TOO MUCH SPEND, TOO LITTLE BENEFIT?



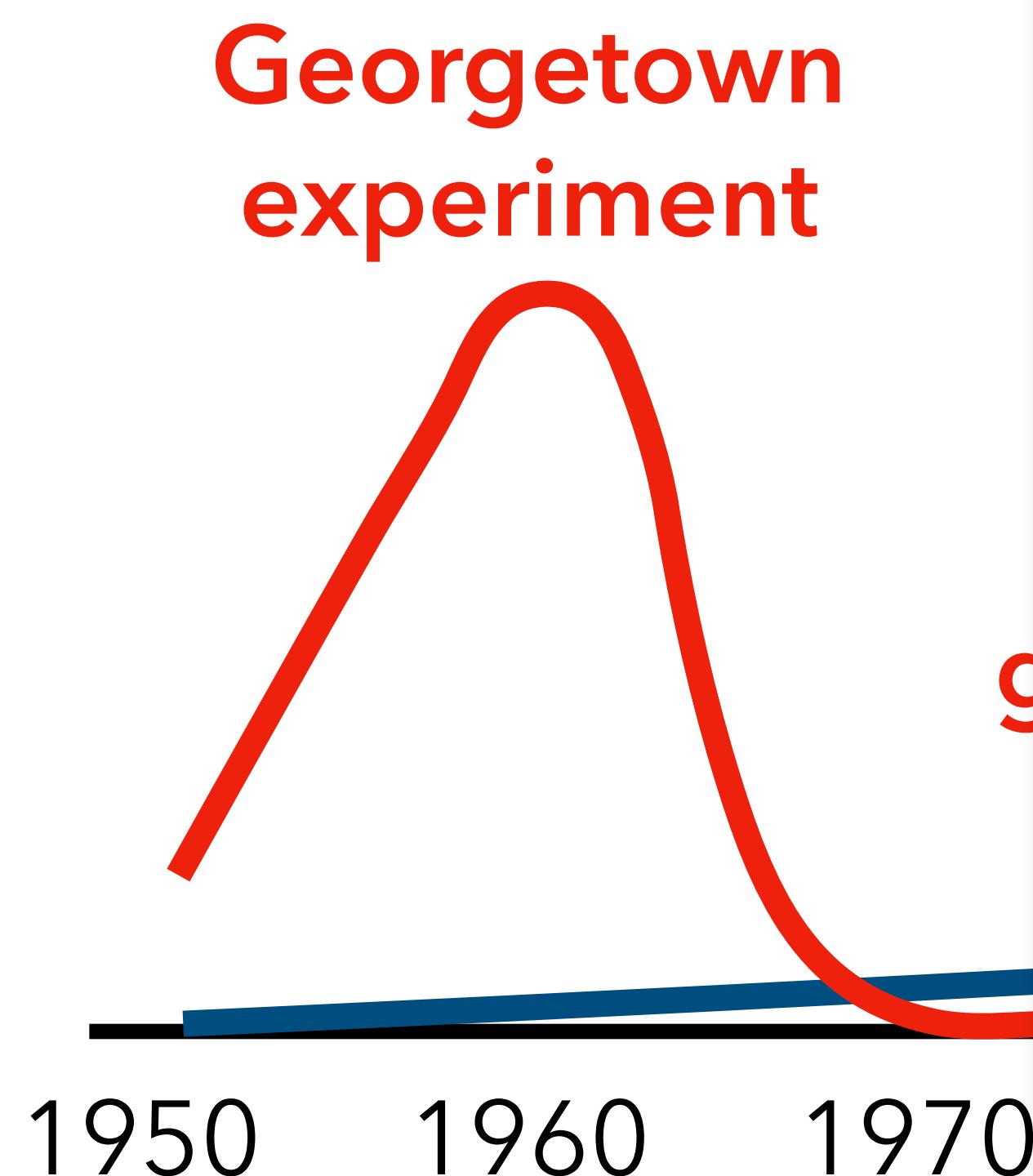
Allison Nathan: You haven't bought into the current generative AI enthusiasm nearly as much as many others. Why is that?

Jim Covello: My main concern is that the substantial cost to develop and run AI technology means that AI applications must solve extremely complex and important problems for enterprises to earn an appropriate return on investment (ROI). We estimate that the AI infrastructure buildout will cost over \$1tn in the next several years alone, which includes spending on data centers, utilities, and applications. So, the crucial question is: What \$1tn problem will AI solve? Replacing low-wage jobs with tremendously costly technology is basically the polar opposite of the prior technology transitions I've witnessed in my thirty years of closely following the tech industry.

**Goldman Sachs, 25 juin 2024**



# Hype vs. reality

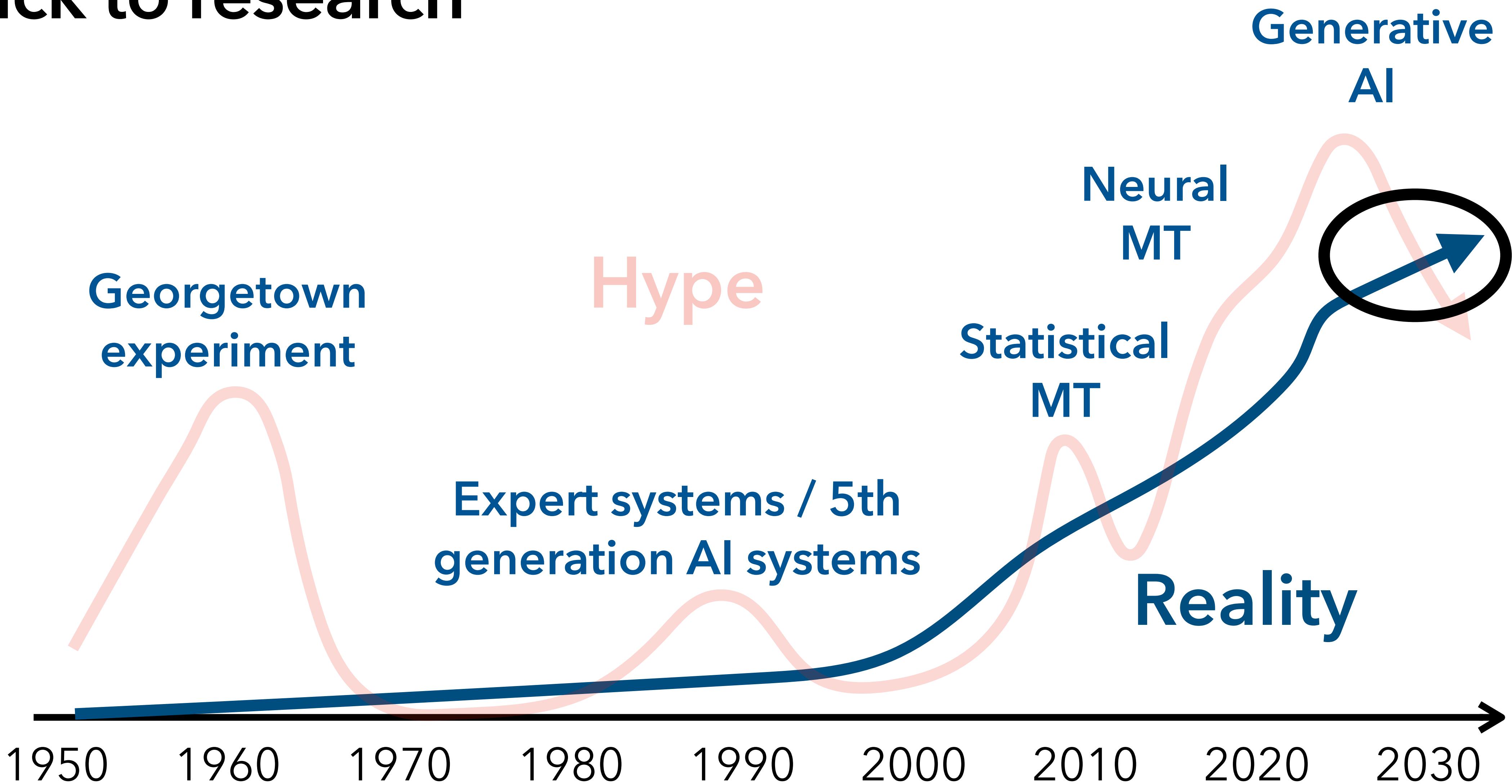


The screenshot shows a news article from FORTUNE magazine. The headline reads: "Alibaba's chair issues 'bubble' warning about U.S. AI investments". The author is listed as DAVID NEYER, with a publication date of March 25, 2025 at 4:01 PM (GMT+1). Below the headline is a photograph of Joseph Tsai, wearing a black Nike hoodie with a teal logo, standing outdoors with his hands clasped. A caption at the bottom states: "Alibaba Group chair and New York Liberty owner Joseph Tsai, pictured during the championship ticker tape parade in October 2024." The source is ELSA—GETTY IMAGES.



Figure adapted from (Koehn 2017) and speculatively extended beyond today

# Back to research





A detailed steampunk illustration featuring a massive, multi-tiered clock tower that spirals upwards like a helix. The tower is intricately decorated with gears, pipes, and mechanical components, with steam rising from various points. In the foreground, a large industrial complex with conveyor belts and machinery is visible. The sky above is filled with numerous airships, dirigibles, and small flying vehicles, set against a backdrop of a vast landscape and a cloudy sky.

Questions?