Why Language Models Hallucinate

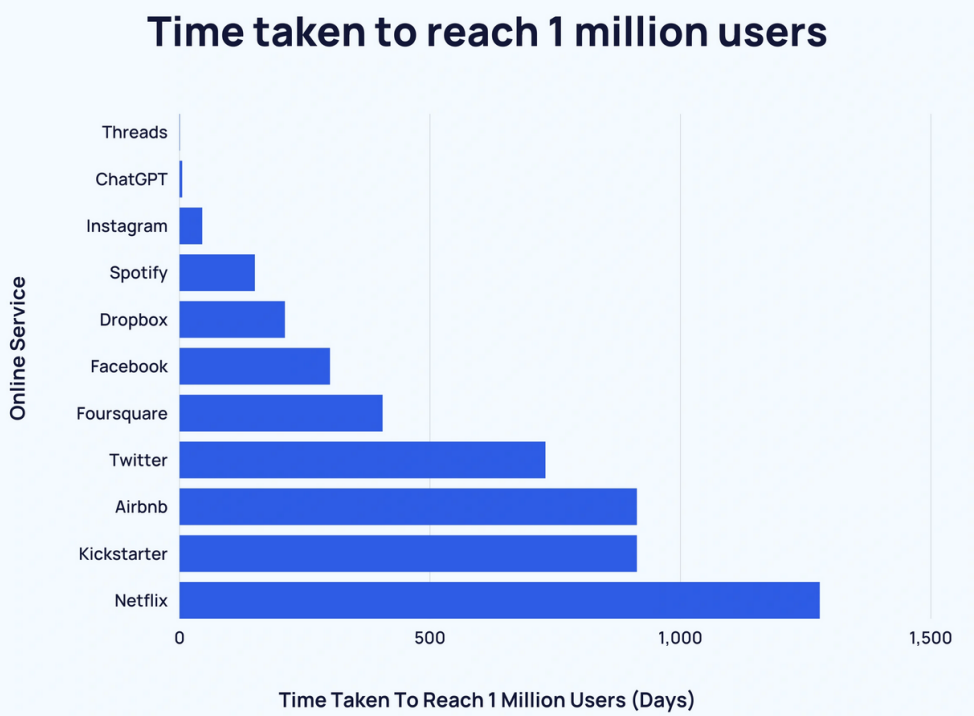
[Kalai et al]

Martin Nemjo

Agenda

- Large language model phenomenon
- Problem with large language models: hallucinations
- Intermezzo: How LLMs are trained
- Related work to hallucinations
- Paper approaches

Large language model phenomenon



Attention Is All You Need

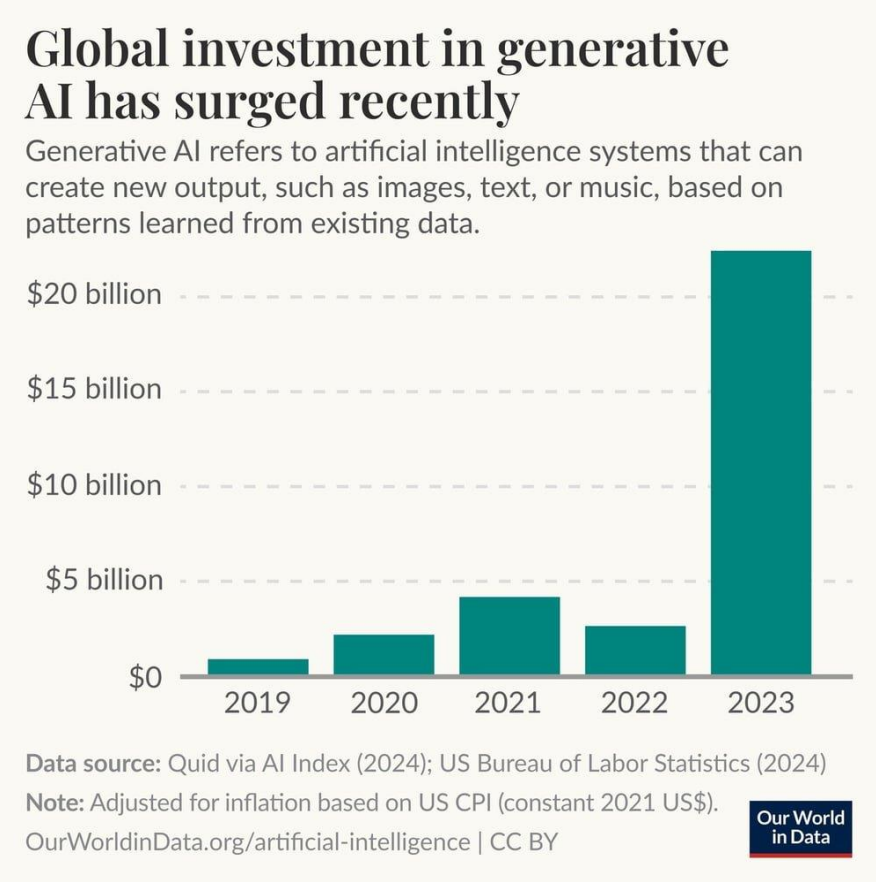
Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Gemini

Large language model phenomenon



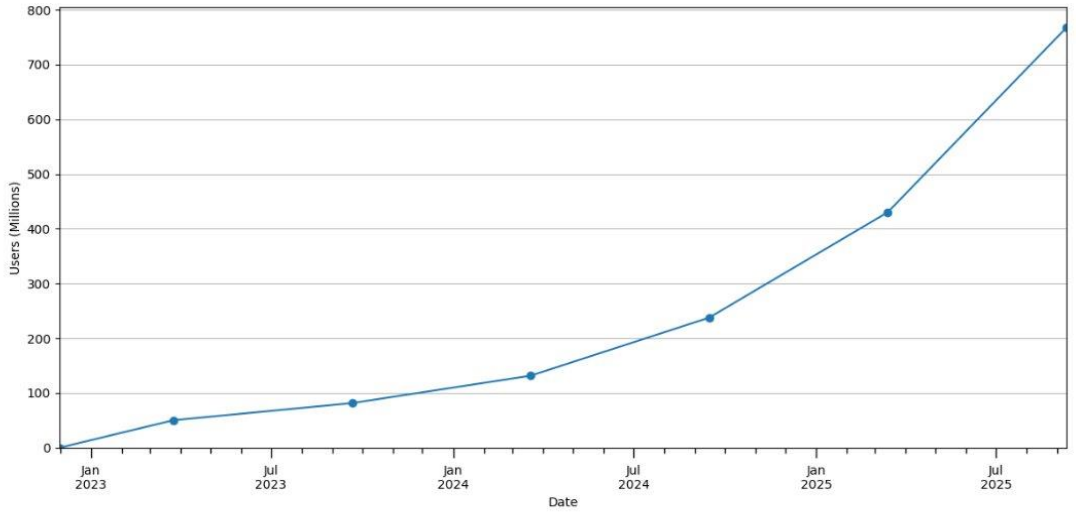
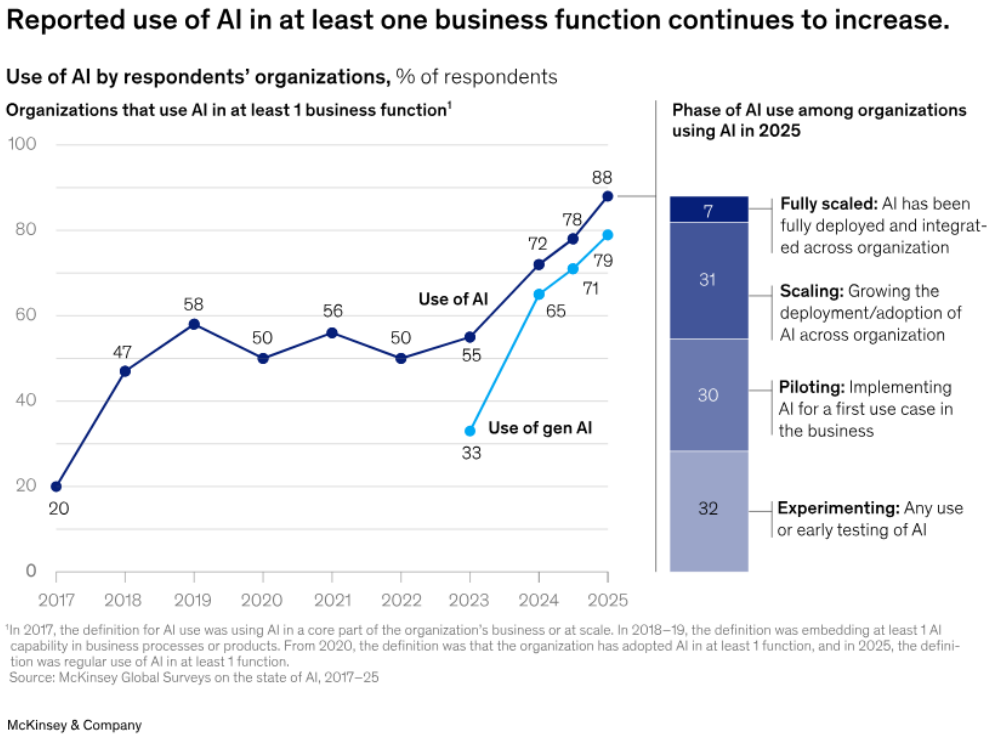
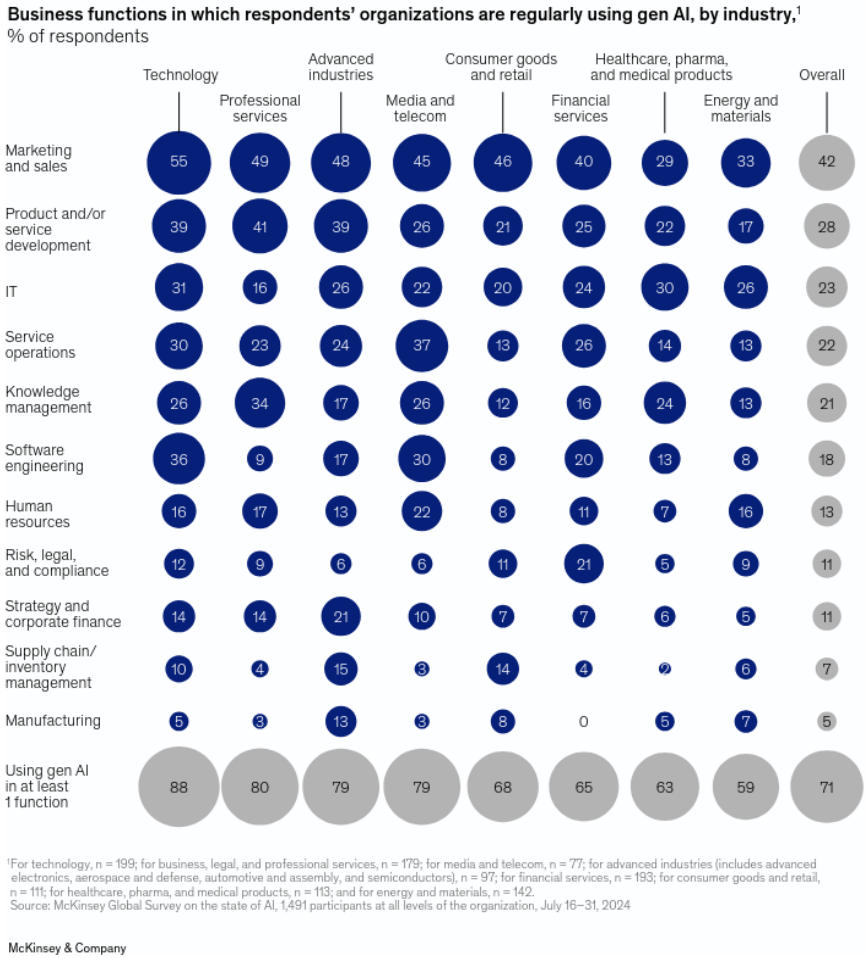


Figure 3: Weekly active ChatGPT users on consumer plans (Free, Plus, Pro), shown as point-in-time snapshots every six months, November 2022–September 2025.

LLM usage in business





But..

BBC

Home News Sport Business Technology Health Culture Arts Travel Earth Audio Video Live

ChatGPT: US lawyer admits using AI for case research

28 May 2023

Share Save

Kathryn Armstrong

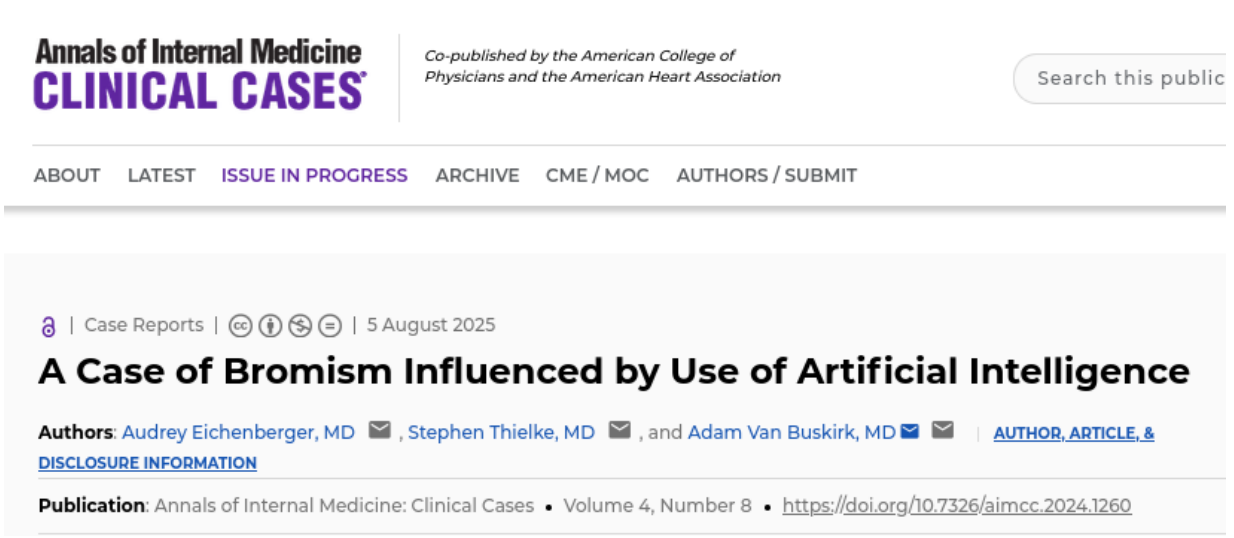
BBC News

Reuters

US lawyer sanctioned after being caught using ChatGPT for court brief

Richard Bednar apologized after Utah appeals court discovered false citations, including one nonexistent case

And..



Annals of Internal Medicine CLINICAL CASES

Co-published by the American College of Physicians and the American Heart Association

Search this public

ABOUT LATEST ISSUE IN PROGRESS ARCHIVE CME / MOC AUTHORS / SUBMIT

Case Reports | 5 August 2025

A Case of Bromism Influenced by Use of Artificial Intelligence

Authors: Audrey Eichenberger, MD , Stephen Thielke, MD , and Adam Van Buskirk, MD AUTHOR, ARTICLE, & DISCLOSURE INFORMATION

Publication: Annals of Internal Medicine: Clinical Cases • Volume 4, Number 8 • https://doi.org/10.7326/aimcc.2024.1260

"You're Not Crazy": A Case of New-onset AI-associated Psychosis

Due to the timely nature of this topic, we are providing an advanced release of this article, ahead of the October-December 2025 issue publication. This article is subject to changes following final review from the authors and editorial staff.

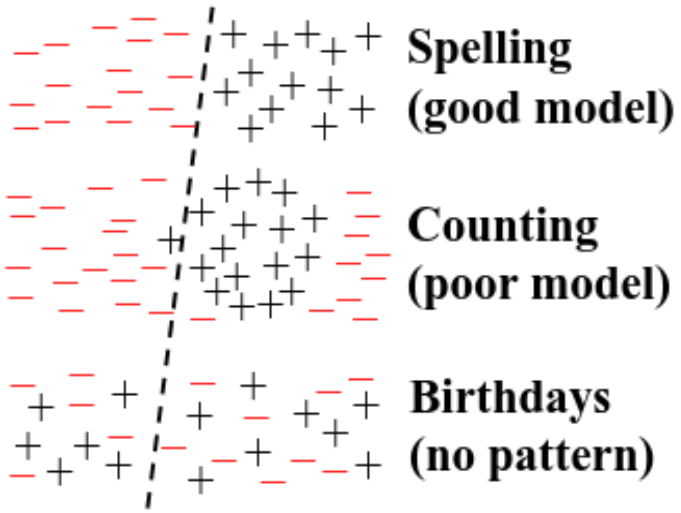
Innov Clin Neurosci. 2025;22(10–12). Epub ahead of print.

by Joseph M. Pierre, MD; Ben Gaeta, MD; Govind Raghavan, MD; and Karthik V. Sarma, MD, PhD

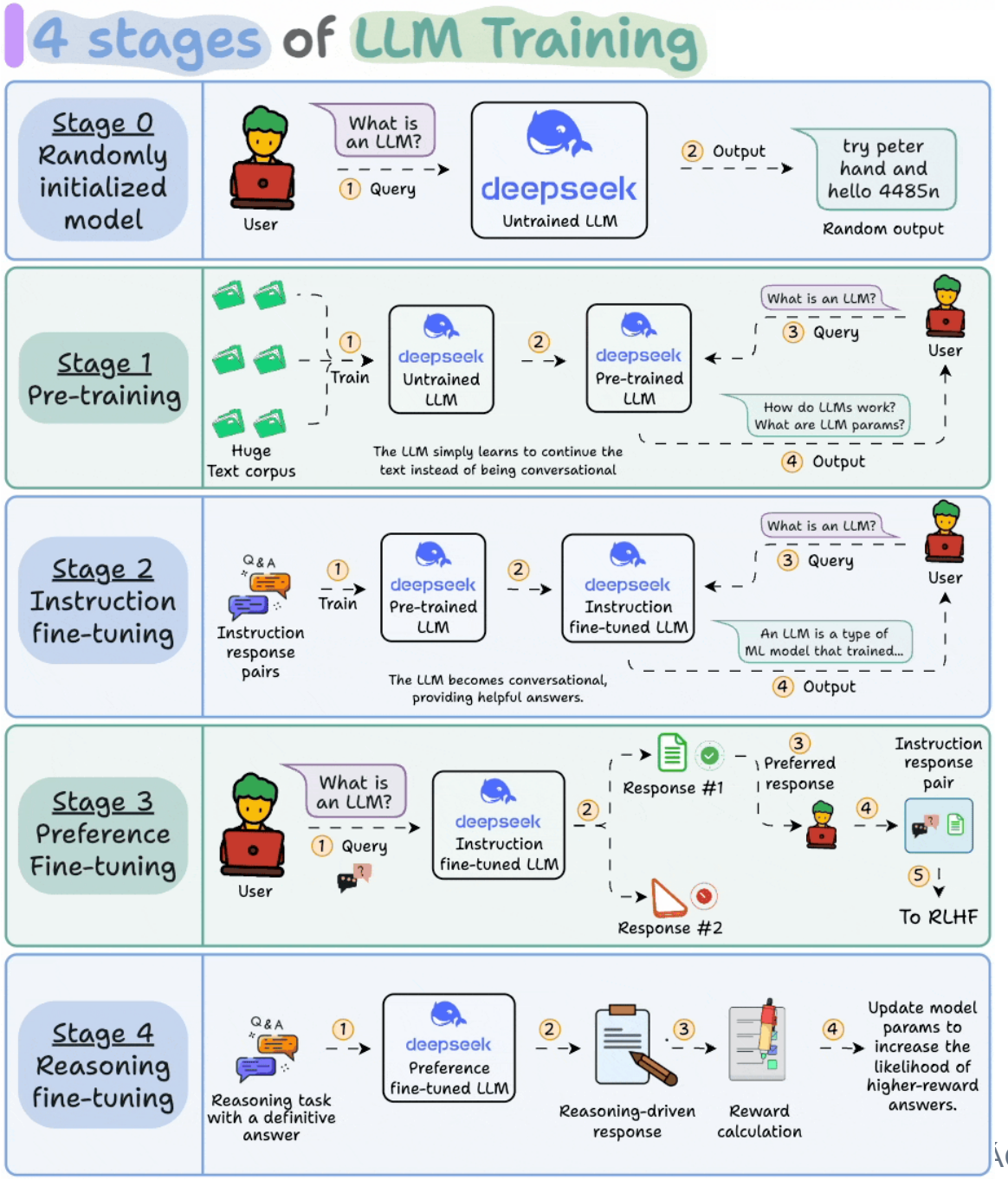
All authors are with the University of California, San Francisco in San Francisco, California.

Hallucinations as described in the paper

Valid examples +	**Error examples** –
Greetings.	Greatings.
How can I help?	How kan eye help?
There are 2 D's in LADDER.	There are 3 L's in SPELL.
There is 1 N in PIANO.	There is 1 G in CAT.
Mia Holdner's birthday is 4/1.	Colin Merivale's birthday is 8/29.
I don't know Zdan's birthday.	Jago Pere's birthday is 8/21.

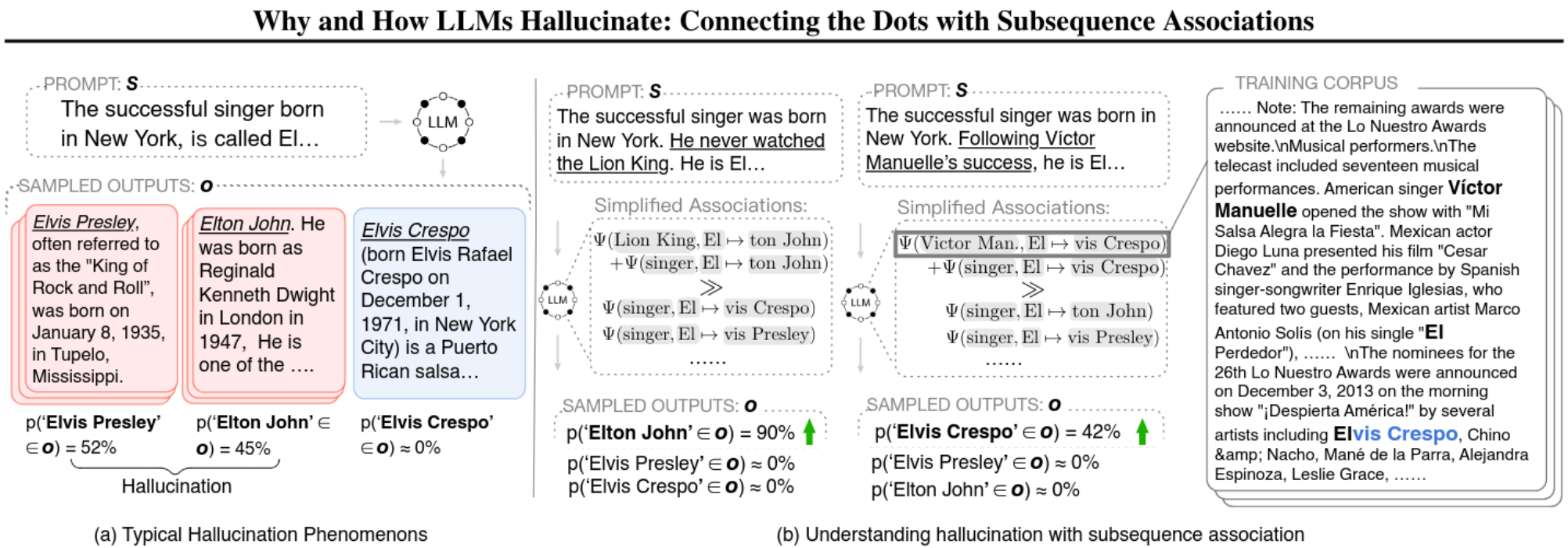


ϵ is a general set of errors, V is a set of plausible (valid) responses and $X = V \cup \epsilon$ is the set of plausbile output strings



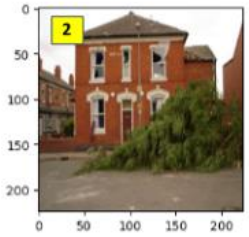
How LLMs are trained

Hallucinations approaches: subsequence associations

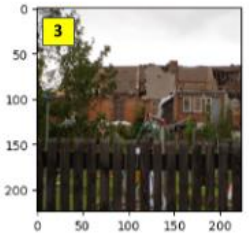


Hallucinations approaches: context hijacking

There was an emergency situation after the storm.



Many trees had fallen and, a building was damaged.



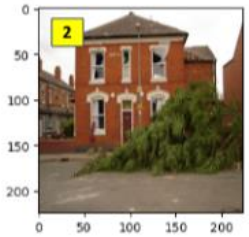
The damage to the building was very severe.

Cars nearby were damaged as well.

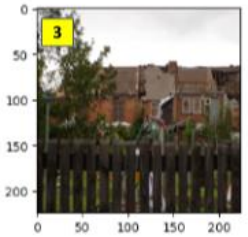
Coherent response

: person was walking his dog when he heard a loud bang and saw a tree had fallen on a house.

There was an emergency situation after the storm.

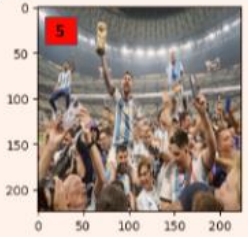


Many trees had fallen and, a building was damaged.



The damage to the building was very severe.

Cars nearby were damaged as well.



Lionel Messi holds up the World Cup trophy after Argentina defeated France in the tournament final.

Incoherent & hijacked response

: person is a fan of the game.

Paper approach: IIV classifier distribution

We now formalize the IIV binary-classification problem, introduced in the introduction. IIV is specified by the target function $f : \mathcal{X} \to \{-, +\}$ to be learned (membership in $\mathcal{V}$) and the distribution D over examples $\mathcal{X}$ (a 50/50 mix of samples from p and uniformly random errors):

$$D(x) := \begin{cases} p(x)/2 & \text{if } x \in \mathcal{V}, \\ 1/2|\mathcal{E}| & \text{if } x \in \mathcal{E}, \end{cases} \text{ and } f(x) := \begin{cases} + & \text{if } x \in \mathcal{V}, \\ - & \text{if } x \in \mathcal{E}. \end{cases}$$

Our analysis lower bounds the error rate $\mathrm{err} = \hat{p}(\mathcal{E})$ in terms of IIV's aforementioned *misclassification rate* $\mathrm{err}_{\mathrm{iiv}}$:

$$\mathrm{err}_{\mathrm{iiv}} := \Pr_{x \sim D}\left[\hat{f}(x) \neq f(x)\right], \text{ where } \hat{f}(x) := \begin{cases} + & \text{if } \hat{p}(x) > 1/|\mathcal{E}|, \\ - & \text{if } \hat{p}(x) \leq 1/|\mathcal{E}|. \end{cases} \quad (2)$$

Paper: main result

$$\mathrm{err} := \hat{p}(\mathcal{E}) = \Pr_{x \sim \hat{p}}[x \in \mathcal{E}]. \tag{1}$$

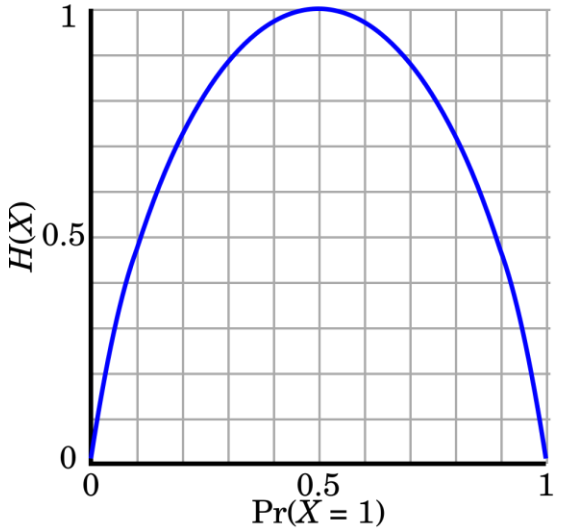
Corollary 1. *For any training distribution p such that $p(\mathcal{V}) = 1$ and any base model $\hat{p}$,*

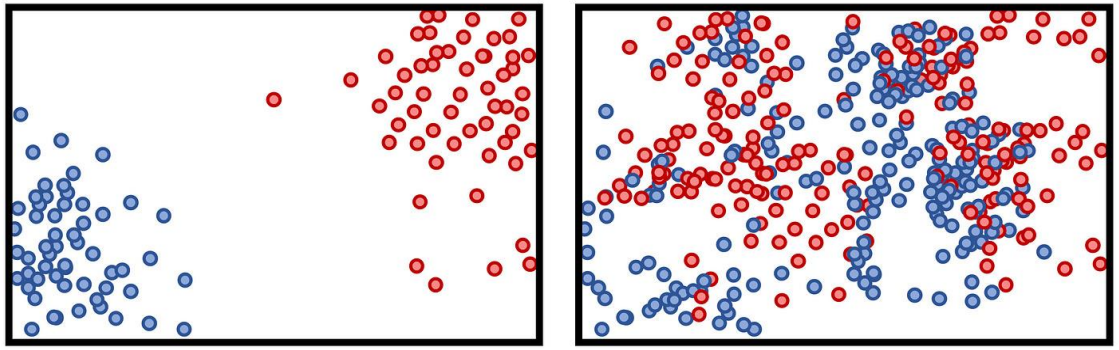
$$\mathrm{err} \geq 2 \cdot \mathrm{err}_{\mathrm{iiv}} - \frac{|\mathcal{V}|}{|\mathcal{E}|} - \delta,$$

for $\mathrm{err}, \mathrm{err}_{\mathrm{iiv}}$ from Eqs. (1) and (2), and $\delta := |\hat{p}(\mathcal{A}) - p(\mathcal{A})|$ for $\mathcal{A} := \{x \in \mathcal{X} \mid \hat{p}(x) > 1/|\mathcal{E}|\}$.

Entropy

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x),$$



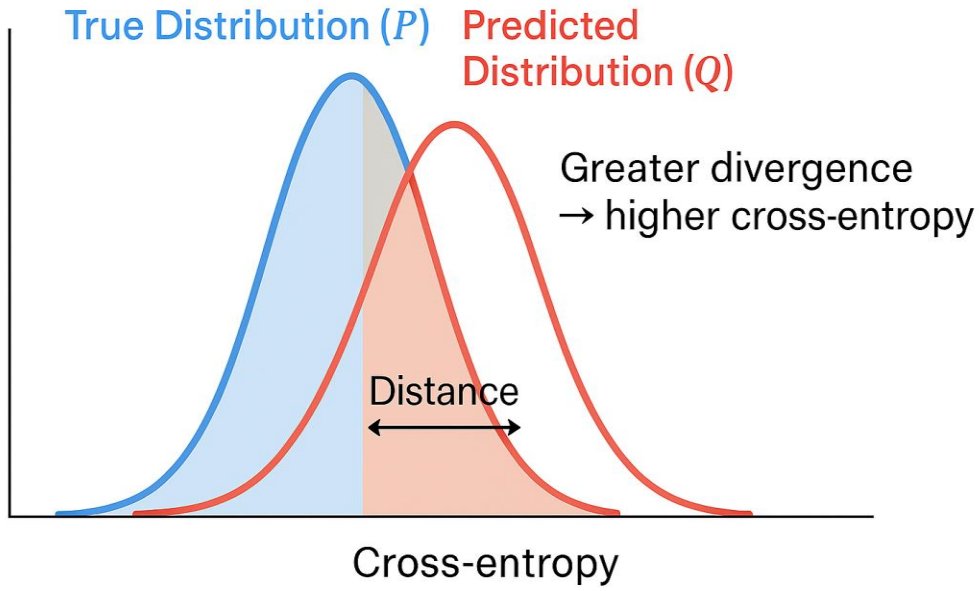


Low Entropy

High Entropy

Cross-entropy

$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x).$$



Main result: calibration

Here is a particularly simple justification for why δ is typically small for the standard pretraining cross-entropy objective,

$$\mathcal{L}(\hat{p}) = \mathop{\mathbb{E}}_{x\sim p}[-\log \hat{p}(x)]. \tag{3}$$

Consider rescaling the probabilities of the positively-labeled examples by a factor $s > 0$ and normalizing:

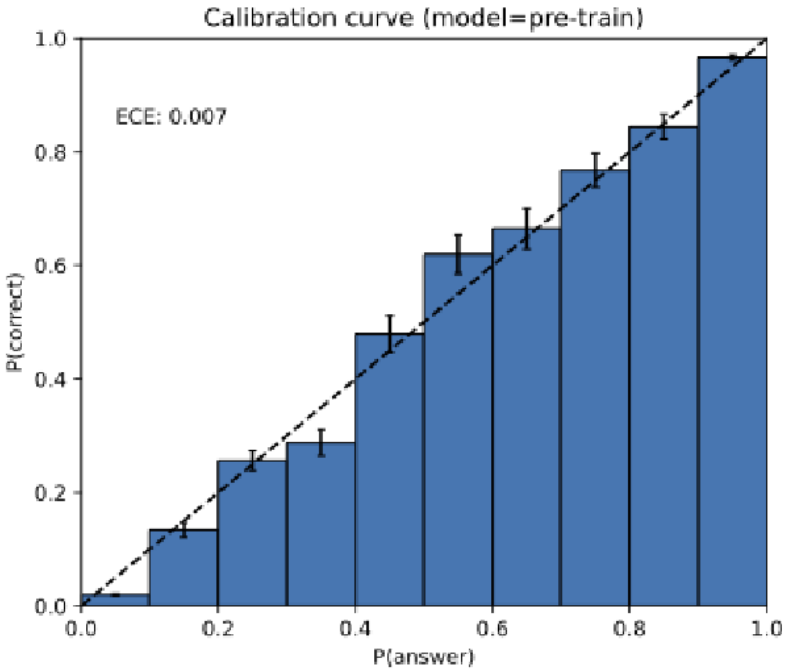
$$\hat{p}_s(x) :\propto \begin{cases} s \cdot \hat{p}(x) & \text{if } \hat{p}(x) > 1/|\mathcal{E}|, \\ \hat{p}(x) & \text{if } \hat{p}(x) \leq 1/|\mathcal{E}|. \end{cases}$$

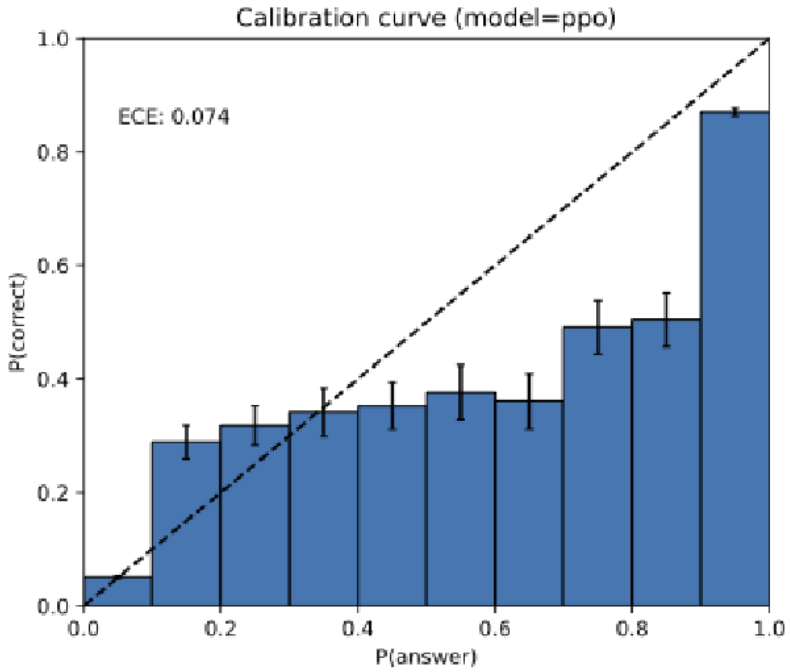
Then, a simple calculation shows that δ is the magnitude of the derivative of the loss with respect to the scaling factor s, evaluated at $s = 1$:

$$\delta = \left| \frac{d}{ds}\mathcal{L}(\hat{p}_s)\Big|_{s=1} \right|.$$

If $\delta \neq 0$, then rescaling by some $s \neq 1$ would reduce the loss, so the loss is not at a local minimum.

Example: ChatGPT pre-training calibration





What about prompts?

Theorem 1. *For any training distribution p such that $p(\mathcal{V}) = 1$ and any base model $\hat{p}$,*

$$\mathrm{err} \geq 2 \cdot \mathrm{err}_{\mathrm{iiv}} - \frac{\max_c |\mathcal{V}_c|}{\min_c |\mathcal{E}_c|} - \delta,$$

where $\delta := |\hat{p}(\mathcal{A}) - p(\mathcal{A})|$ *for* $\mathcal{A} := \{(c, r) \in \mathcal{X} \mid \hat{p}(r \mid c) > 1/\min_c |\mathcal{E}_c|\}$.

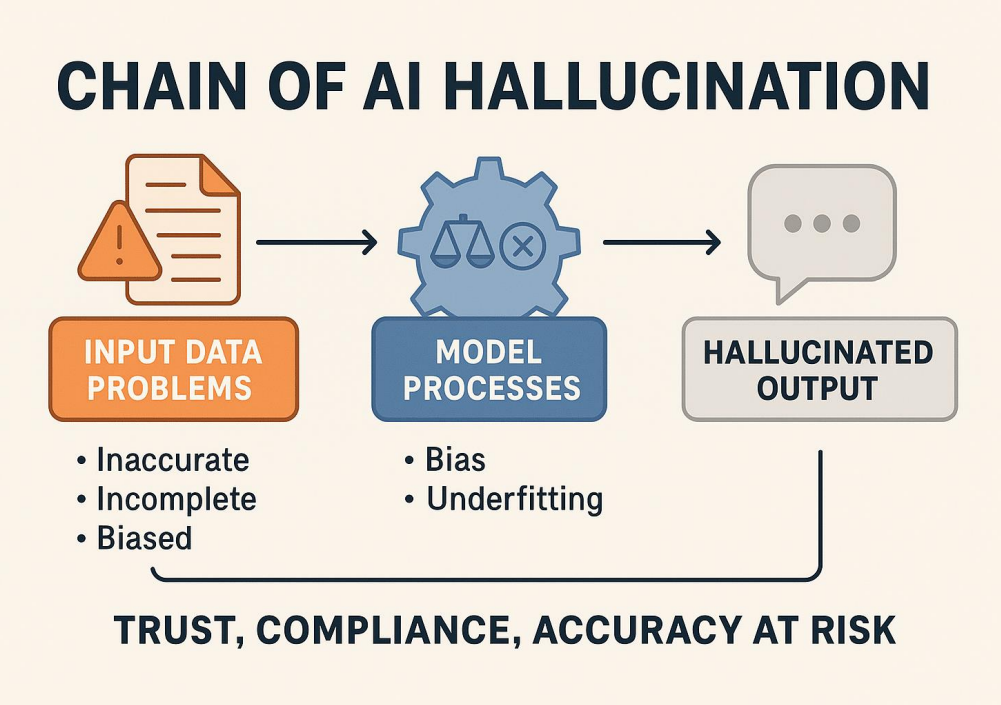
Additional factors for errors: simple models

$$\mathrm{opt}(\mathcal{G}) := \min_{g \in \mathcal{G}} \Pr_{x \sim D}[g(x) \neq f(x)] \in [0,1]. \qquad \mathrm{err} \geq 2 \cdot \mathrm{opt}(\mathcal{G}) - \frac{\max_c |\mathcal{V}_c|}{\min_c |\mathcal{E}_c|} - \delta.$$

Theorem 3 (Pure multiple-choice). *Suppose* $|\mathcal{V}_c| = 1$ *for all* $c \in \mathcal{C}$ *and let* $C = \min_c |\mathcal{E}_c| + 1$ *be the number of choices. Then,*

$$\mathrm{err} \geq 2\left(1 - \frac{1}{C}\right) \cdot \mathrm{opt}(\mathcal{G})$$

Additional factors for errors



Computational hardness (laws of computational complexity)

Distribution shift (out-of-distribution prompts)

Garbage in, Garbage out

Post-training errors

Benchmark	Scoring method	Binary grading	IDK credit
GPQA	Multiple-choice accuracy	Yes	None
MMLU-Pro	Multiple-choice accuracy	Yes	None
IFEval	Programmatic instruction verification	Yes[a]	None
Omni-MATH	Equivalence grading*	Yes	None
WildBench	LM-graded rubric*	No	Partial[b]
BBH	Multiple-choice / exact-match	Yes	None
MATH (L5 split)	Equivalence grading*	Yes	None
MuSR	Multiple-choice accuracy	Yes	None
SWE-bench	Patch passes unit tests	Yes	None
HLE	Multiple-choice / equivalence grading*	Yes	None

Proposed approach

Answer only if you are $> t$ confident, since mistakes are penalized $t/(1-t)$ points, while correct answers receive 1 point, and an answer of "I don't know" receives 0 points.

Thank you for your attention

Martin Nemjo