Why Language Models Hallucinate

Adam Tauman Kalai* OpenAI

Ofir Nachum OpenAI

Santosh S. Vempala† Georgia Tech

Edwin Zhang OpenAI

September 4, 2025

陈佳含 2025.09.16

01 Introduction

What was the title of Adam Kalai's dissertation?

GPT-4o	Adam Tauman Kalai's Ph.D. dissertation (completed in 2002 at CMU) is entitled: "Boosting, Online Algorithms, and Other Topics in Machine Learning."
DeepSeek	Algebraic Methods in Interactive Machine Learning". . . at Harvard University in 2005.
Llama	Efficient Algorithms for Learning and Playing Games". . . in 2007 at MIT.

Hallucinations are an important special case of errors produced by language models.

Plausible falsehoods called hallucinations.

01 Introduction

Even if the training data were error-free, the objectives optimized during language model training would lead to errors being generated.

- **Errors caused by pretraining**
- **Why hallucinations survive post-training**

02 Pretrain

The key challenge in proving that **base models(after pretraining)** err is that many language models do not err.

- The degenerate model which always outputs IDK.
- Assuming error-free training data, the trivial base model which regurgitates text from a random training example.
- The optimal base model, but prohibitively large training data.

---> well-trained base models <--

02 Pretrain - Reduction without Prompts

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

FOCUS: Generating valid outputs is harder than classifying output validity.

Without prompts, a base model $\hat{p}$ is a probability distribution over a set X (X = E ∪ V) .

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

IIV：

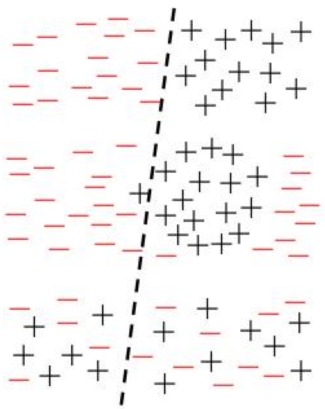
Target function： $f : \mathcal{X} \to \{-, +\}$

Distribution D：

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

02 Pretrain - Reduction without Prompts

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.



IIV：

Target function： $f : \mathcal{X} \rightarrow \{-,+\}$

Distribution D：

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

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

The misclassification rate：

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

Corollary 1：

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

$$\delta := |\hat{p}(\mathcal{A}) - p(\mathcal{A})| \ \mathit{for}\ \mathcal{A} := \{x \in \mathcal{X} \mid \hat{p}(x) > 1/|\mathcal{E}|\}$$

02 Pretrain - Reduction without Prompts

The standard pretraining cross-entropy objective:

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

Rescale the probabilities of the positively-labeled examples:

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

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

02 Pretrain - Reduction without Prompts

The misclassification rate：

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

Corollary 1：

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

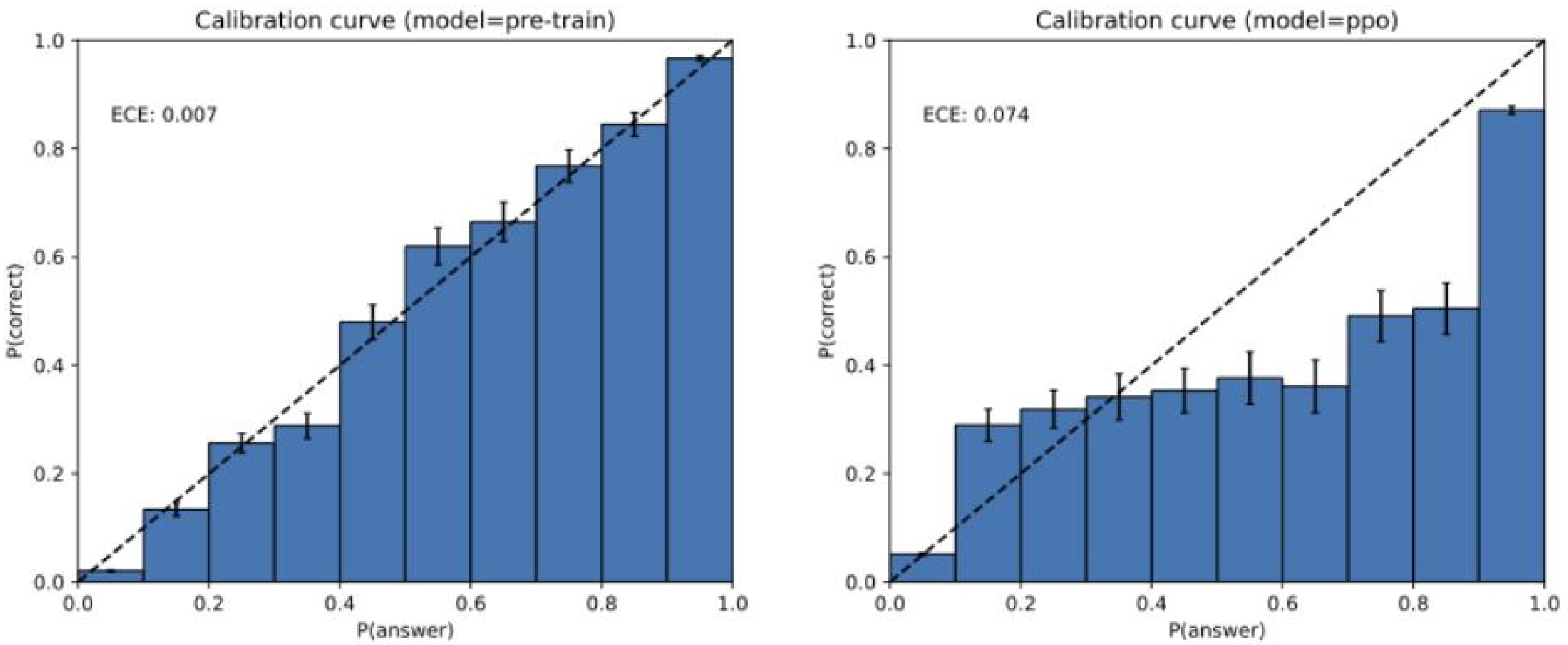
$$\delta := |\hat{p}(\mathcal{A}) - p(\mathcal{A})| \ \mathit{for}\ \mathcal{A} := \{x \in \mathcal{X} \mid \hat{p}(x) > 1/|\mathcal{E}|\}$$

$$\boxed{\mathrm{err}_{\mathrm{iiv}} \lesssim 1/2}$$

02 Pretrain - Reduction without Prompts

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

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



GPT-4 calibration histograms before (left) and after (right) reinforcement learning

02 Pretrain - Reduction with Prompts

FOCUS: Generating valid outputs is harder than classifying output validity.

Without prompts, a base model $\hat{p}$ is a probability distribution over a set X (X = E ∪ V) .

The error rate: $\mathrm{err} := \hat{p}(\mathcal{E}) = \sum_{(c,r)\in\mathcal{E}} \mu(c)\hat{p}(r \mid c)$

IIV：

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

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

02 Pretrain - Arbitrary-fact hallucinations

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Definition 1 (Arbitrary Facts). *The following are fixed: an arbitrary prompt distribution $\mu(c)$, an* IDK *response and, for each prompt c: a response set $\mathcal{R}_c$ and a probability of answering $\alpha_c \in [0,1]$. Independently for each c, a single correct answer $a_c \in \mathcal{R}_c$ is chosen uniformly at random. Finally, $p(a_c \mid c) = \alpha_c$ and $p(\mathrm{IDK} \mid c) = 1 - \alpha_c$ for each $c \in \mathcal{C}$. Thus $\mathcal{E}_c = \mathcal{R}_c \setminus \{a_c\}$ and $\mathcal{V}_c = \{a_c, \mathrm{IDK}\}$.*

Definition 2 (Singleton rate). *A prompt $c \in \mathcal{C}$ is a* singleton *if it appears exactly once in the N training data $\langle (c^{(i)}, r^{(i)}) \rangle_{i=1}^{N}$ without abstention, i.e., $|\{i : c^{(i)} = c \wedge r^{(i)} \neq \mathrm{IDK}\}| = 1$. Let $\mathcal{S} \subseteq \mathcal{C}$ denote the set of singletons and*

$$\mathrm{sr} = \frac{|\mathcal{S}|}{N}$$

denote the fraction of training singletons.

$$1 \longleftarrow \mathrm{err} \geq 2 \cdot \mathrm{err}_{\mathrm{iiv}} - \frac{\max_c |\mathcal{V}_c|}{\min_c |\mathcal{E}_c|} - \delta,$$

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

Alan Turing's estimate of the **unseen-event probability** is the fraction of **samples appearing exactly once**.

02 Pretrain - Pool Models

Misclassifications can also arise when the underlying model is poor because:

- The **model family** cannot represent the concept well.
- The model family is sufficiently expressive but the **model itself is not a good fit**.

02 Pretrain - Pool Models

The minimal error rate of any classifier in a given family G of classifiers g:

$$\mathrm{opt}(\mathcal{G}) := \min_{g \in \mathcal{G}} \Pr_{x \sim D}[g(x) \neq f(x)] \in [0,1]$$

$$\mathcal{G} := \{g_{\theta,t} \mid \theta \in \Theta, t \in [0,1]\}, \text{ where } g_{\theta,t}(c,r) := \begin{cases} + & \text{if } \hat{p}_\theta(r \mid c) > t, \\ - & \text{if } \hat{p}_\theta(r \mid c) \leq t. \end{cases}$$

$$\mathrm{err} \geq 2 \cdot \mathrm{opt}(\mathcal{G}) - \frac{\max_c |\mathcal{V}_c|}{\min_c |\mathcal{E}_c|} - \delta.$$

N-grams model:

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,*

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

e.g.:

c_1 = She lost it and was completely out of... c_2 = He lost it and was completely out of...

r_1 = her mind. r_2 = his mind.

err >= 2*(1-1/2)*(1/2)=0.5

02 Pretrain - Pool Models - Example

DeepSeek-V3 vs DeepSeek-R1:

How many Ds are in DEEPSEEK? If you know, just say the number with no commentary.

DeepSeek-V3	“2” or “3”
DeepSeek-R1	Let me spell it out: D-E-E-P-S-E-E-K. First letter: D — that’s one D. Second letter: E — not D. Third letter: E — not D. . . So, the number of Ds is 1.

02 Pretrain - Additional Factors

Computational Hardness: “What is the decryption of c?”

Distribution shift: OOD.

GIGO: Garbage in, Garbage out.

03 Post-training

How evaluations reinforce hallucination

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

03 Post-training

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.

- Propose making the **confidence threshold explicit** in the instructions.
- Suggest incorporating confidence targets into **existing mainstream evaluations.**

Behavioral calibration： rather than requiring the model to output a probabilistic confidence.