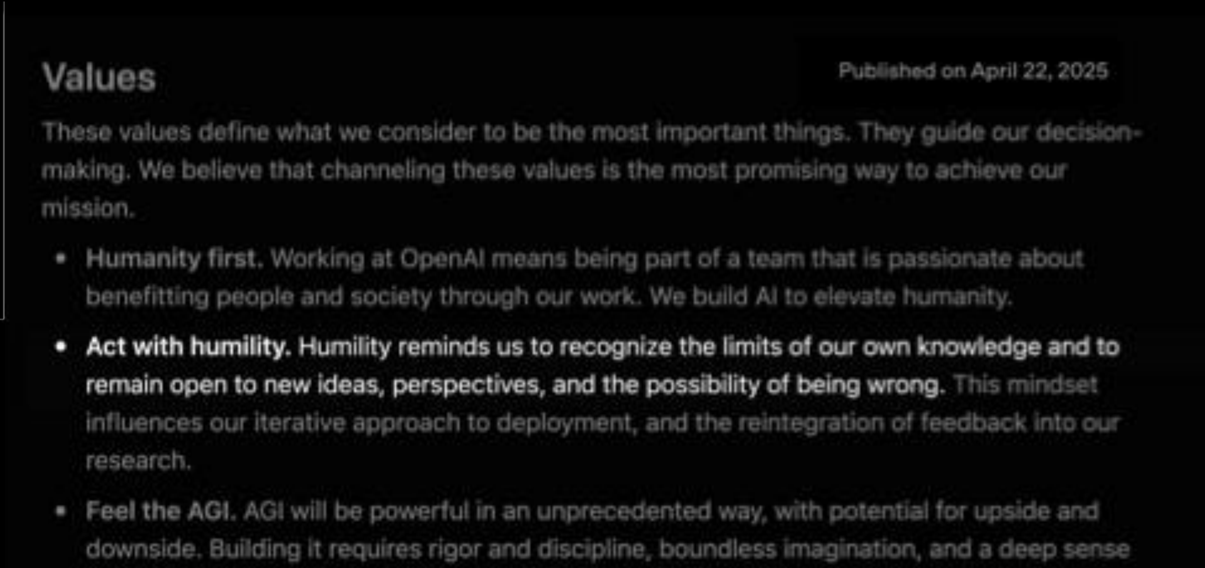
Why Language Models Hallucinate

Adam Tauman Kalai, OpenAI (safety team)
Ofir Nachum, OpenAI
Santosh Vempala, Georgia Tech
Eddie Zhang, OpenAI



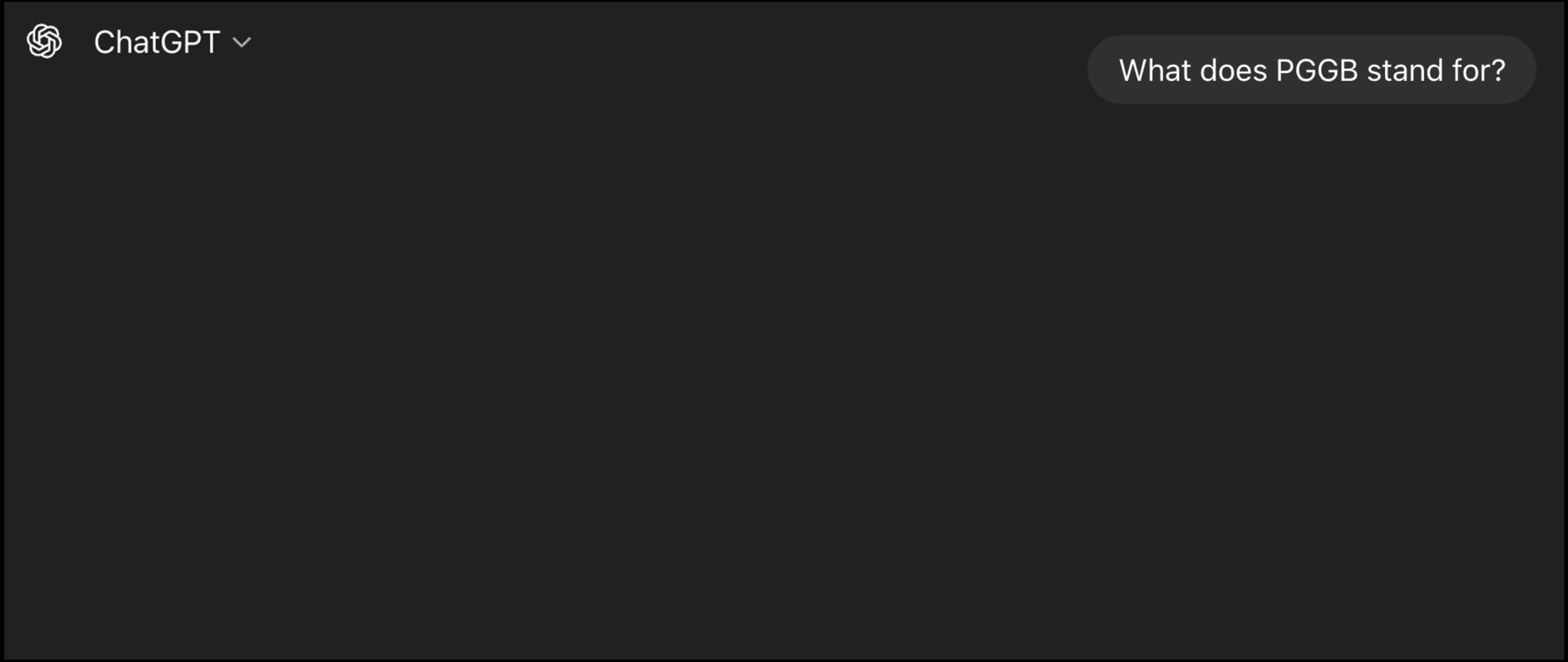
Values

Published on April 22, 2025

These values define what we consider to be the most important things. They guide our decision-making. We believe that channeling these values is the most promising way to achieve our mission.

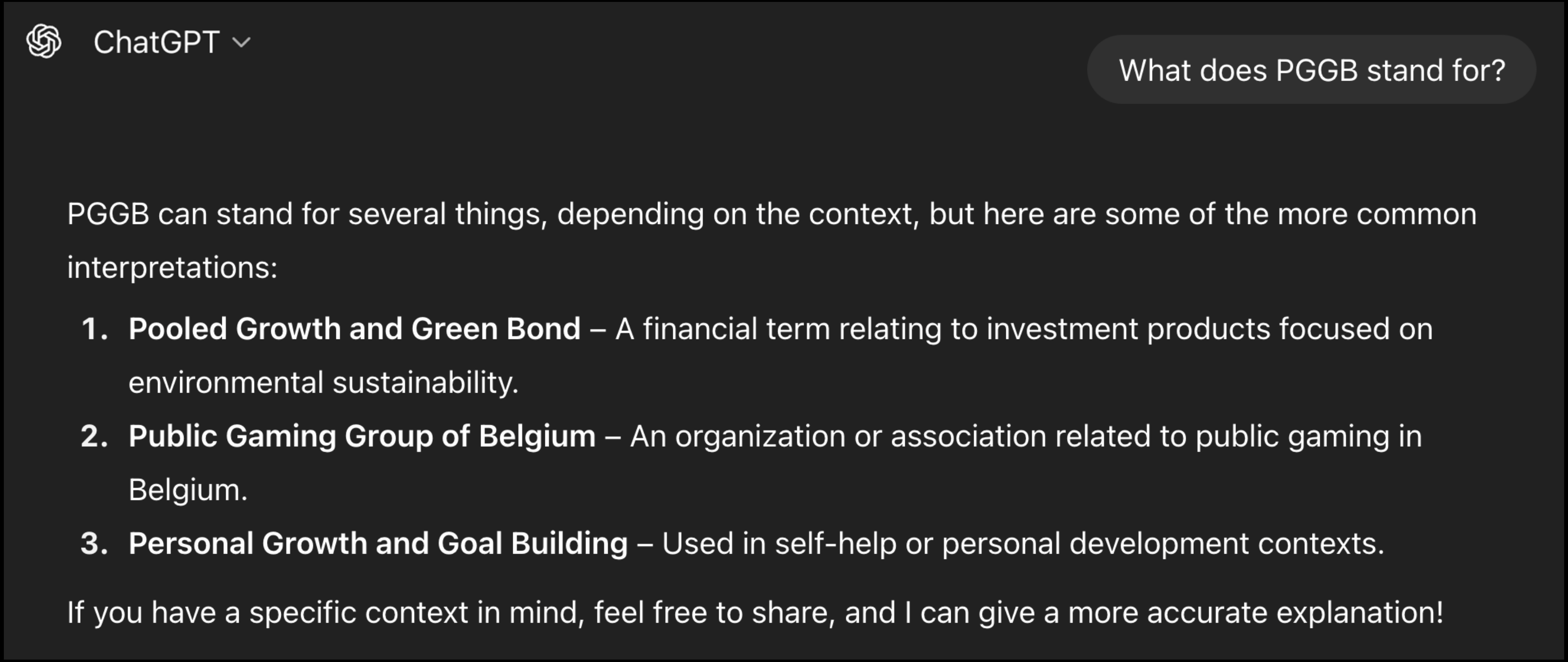
- **Humanity first.** Working at OpenAI means being part of a team that is passionate about benefitting people and society through our work. We build AI to elevate humanity.
- **Act with humility. Humility reminds us to recognize the limits of our own knowledge and to remain open to new ideas, perspectives, and the possibility of being wrong.** This mindset influences our iterative approach to deployment, and the reintegration of feedback into our research.
- **Feel the AGI.** AGI will be powerful in an unprecedented way, with potential for upside and downside. Building it requires rigor and discipline, boundless imagination, and a deep sense

Example



What does PGGB stand for?

Example



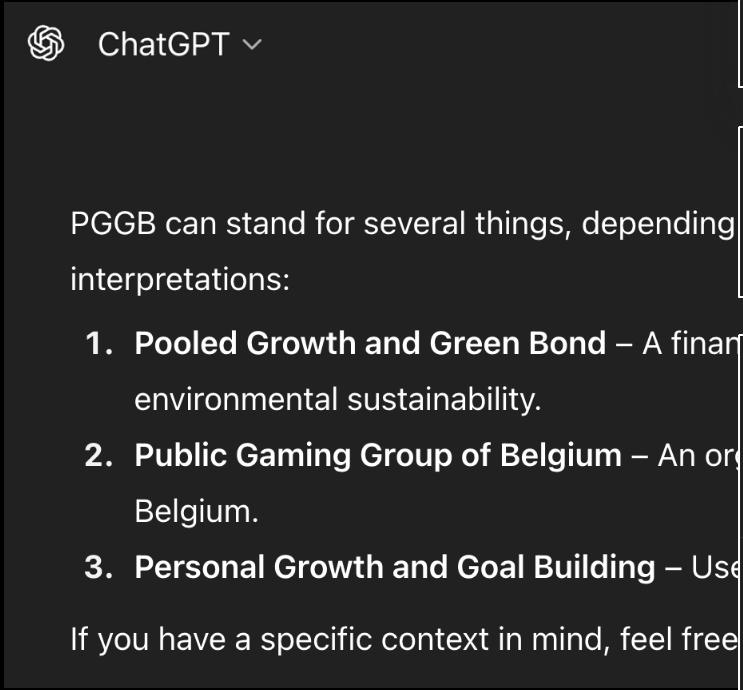
What does PGGB stand for?

PGGB can stand for several things, depending on the context, but here are some of the more common interpretations:

1. **Pooled Growth and Green Bond** – A financial term relating to investment products focused on environmental sustainability.
2. **Public Gaming Group of Belgium** – An organization or association related to public gaming in Belgium.
3. **Personal Growth and Goal Building** – Used in self-help or personal development contexts.

If you have a specific context in mind, feel free to share, and I can give a more accurate explanation!

Example

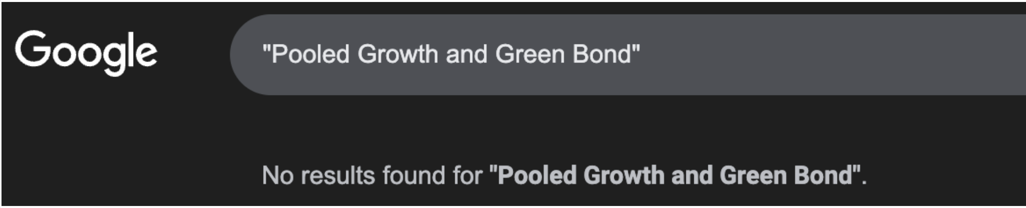


PGGB can stand for several things, depending

interpretations:

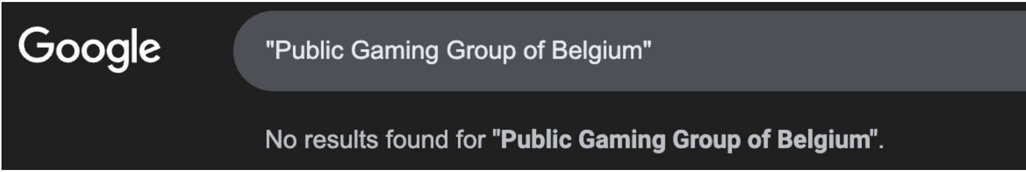
1. **Pooled Growth and Green Bond** – A finan
   environmental sustainability.
2. **Public Gaming Group of Belgium** – An or
   Belgium.
3. **Personal Growth and Goal Building** – Use

If you have a specific context in mind, feel free



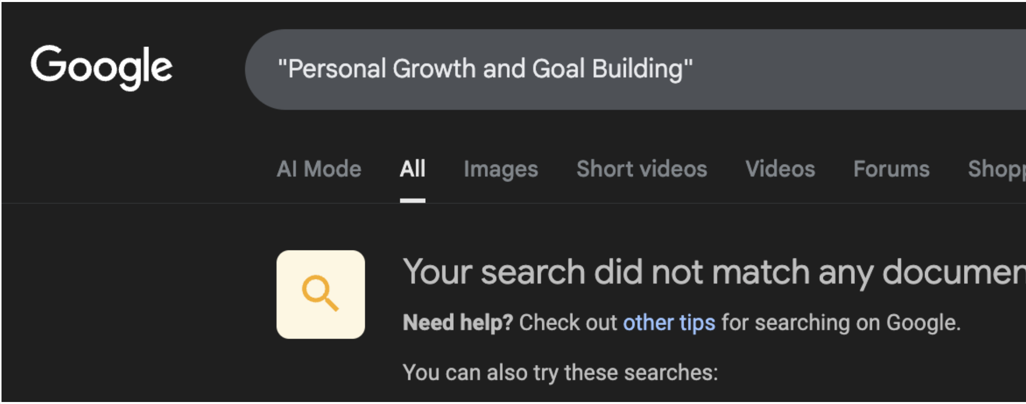
"Pooled Growth and Green Bond"

No results found for **"Pooled Growth and Green Bond"**.



"Public Gaming Group of Belgium"

No results found for **"Public Gaming Group of Belgium"**.

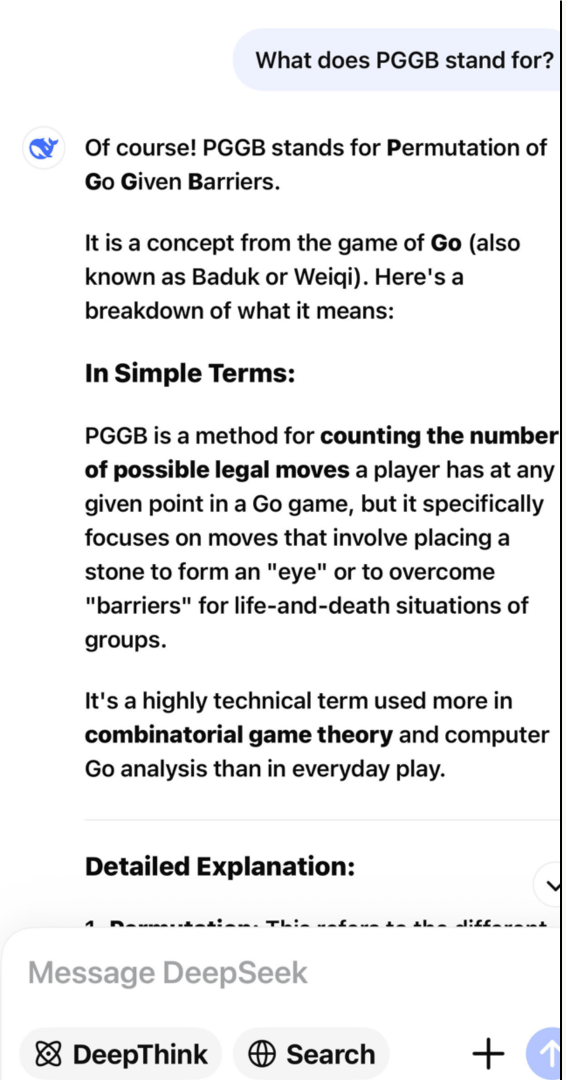


"Personal Growth and Goal Building"

Your search did not match any documen

Need help? Check out other tips for searching on Google.

You can also try these searches:



What does PGGB stand for?

Of course! PGGB stands for **P**ermutation of **G**o **G**iven **B**arriers.

It is a concept from the game of **Go** (also known as Baduk or Weiqi). Here's a breakdown of what it means:

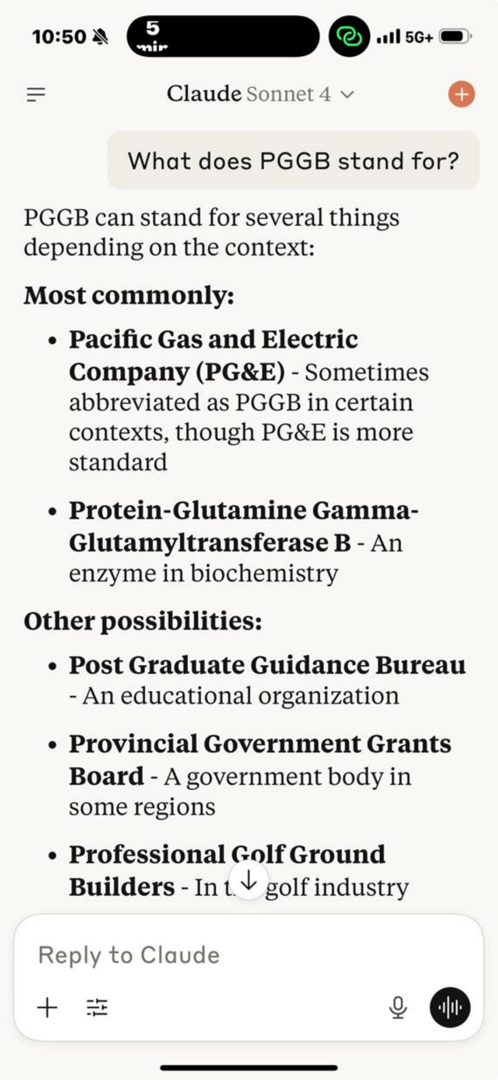
In Simple Terms:

PGGB is a method for **counting the number of possible legal moves** a player has at any given point in a Go game, but it specifically focuses on moves that involve placing a stone to form an "eye" or to overcome "barriers" for life-and-death situations of groups.

It's a highly technical term used more in **combinatorial game theory** and computer Go analysis than in everyday play.

Detailed Explanation:

1. **Permutation:** This refers to the different



What does PGGB stand for?

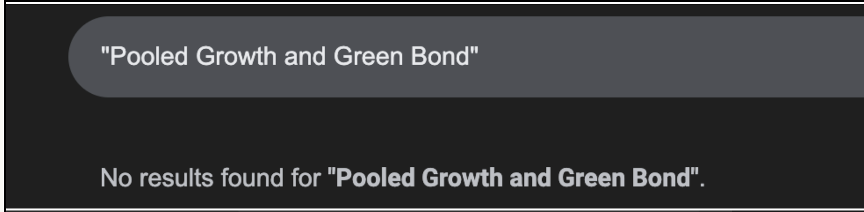
PGGB can stand for several things depending on the context:

Most commonly:

- **Pacific Gas and Electric Company (PG&E)** - Sometimes abbreviated as PGGB in certain contexts, though PG&E is more standard
- **Protein-Glutamine Gamma-Glutamyltransferase B** - An enzyme in biochemistry

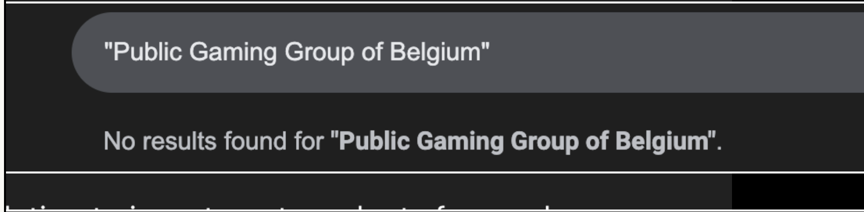
Other possibilities:

- **Post Graduate Guidance Bureau** - An educational organization
- **Provincial Government Grants Board** - A government body in some regions
- **Professional Golf Ground Builders** - In t golf industry



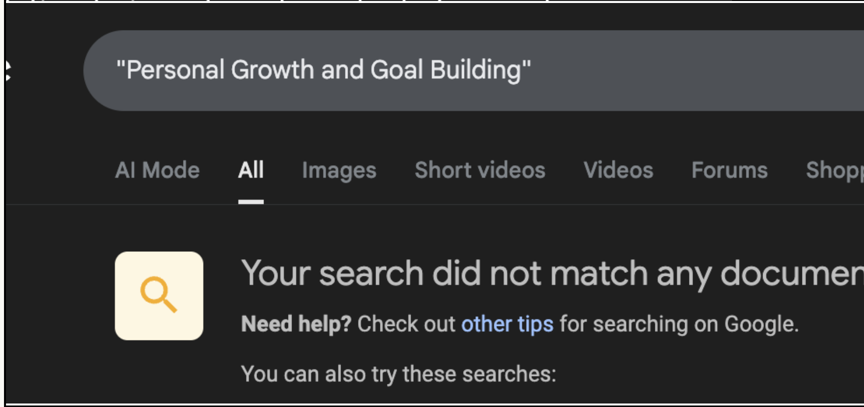
"Pooled Growth and Green Bond"

No results found for **"Pooled Growth and Green Bond"**.



"Public Gaming Group of Belgium"

No results found for **"Public Gaming Group of Belgium"**.



"Personal Growth and Goal Building"

Your search did not match any documen

Need help? Check out other tips for searching on Google.

You can also try these searches:

~~Error~~Word of the Year, Cambridge Dictionary

hallucinate

verb [I or T]

When an artificial intelligence hallucinates, it produces false information:

- LLMs are notorious for hallucinating – generating completely false answers, often supported by fictitious citations.
- The latest version of the chatbot is greatly improved but it will still hallucinate facts.
- Why do generative AI tools hallucinate?

Worse than omission or "I don't know" (humility failure)

🍄

Are you on drugs?

Common misconceptions in the industry

❌ Accuracy measures hallucinations.

$$\text{acc} + \text{idk} + \text{err} = 100\%$$

❌ Humility requires many parameters.

Academia challenge: make a tiny hallucination-free LLM

Did OpenAI just solve hallucinations?

121K views • 8 days ago

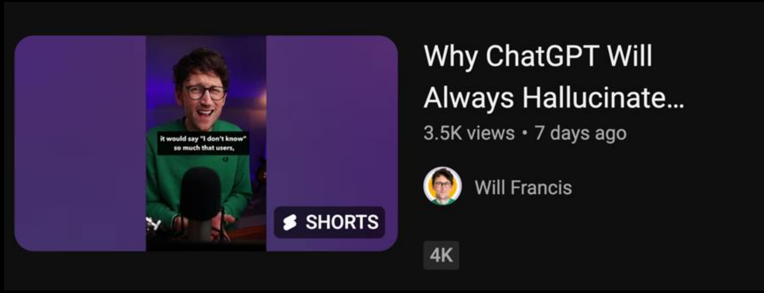
Matthew Berman

❌ Scaling 📈 will lead to 100% accuracy so we don't need humility.

No, there'll always be questions LLMs can't reliably answer.

❌ Hallucinations are inevitable.

No, LLMs should say I don't know.

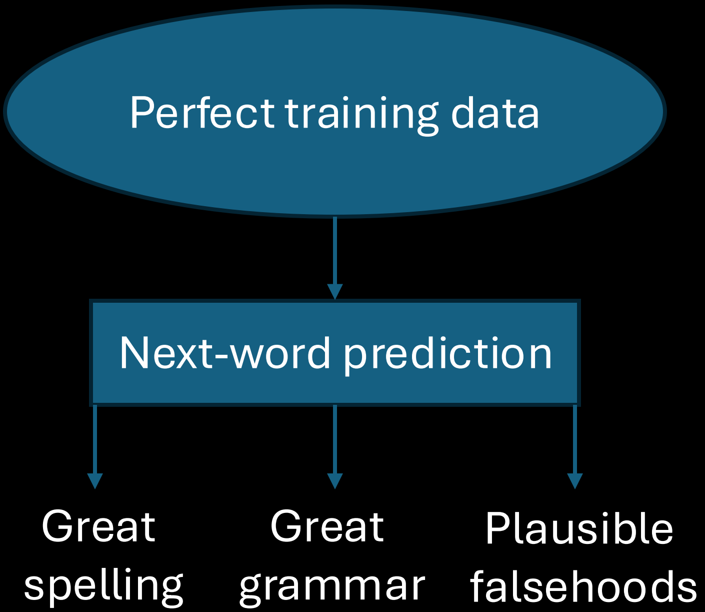


Why ChatGPT Will Always Hallucinate...

3.5K views • 7 days ago

Will Francis

Hallucinations aren’t like other kinds of errors



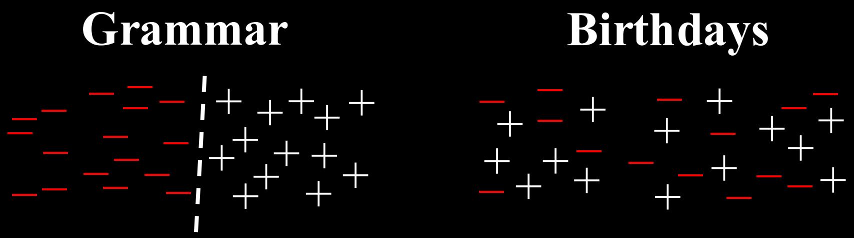
Perfect training data

Next-word prediction

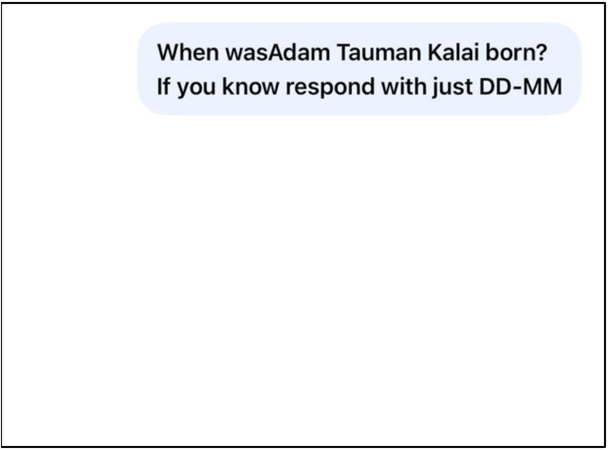
Great spelling

Great grammar

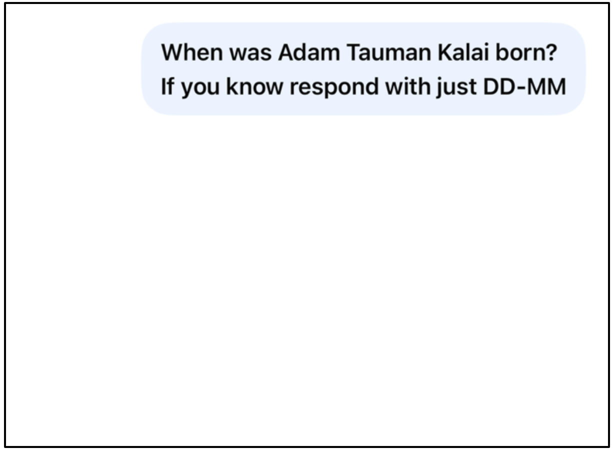
Plausible falsehoods



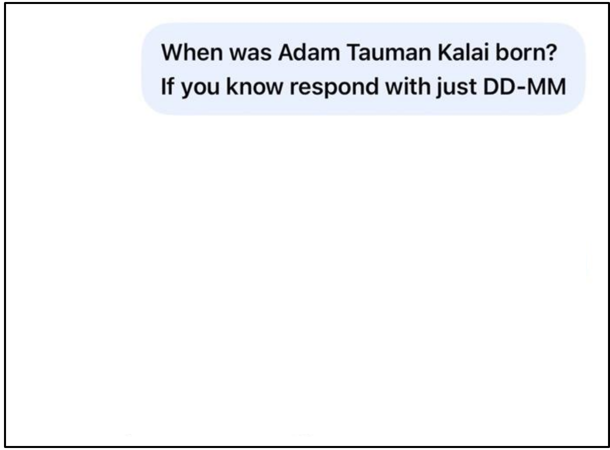
Humility failure: IDK = **“I don’t know”** would often be better



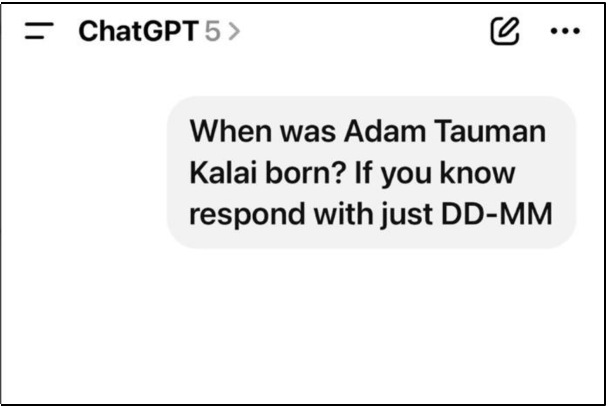
When wasAdam Tauman Kalai born? If you know respond with just DD-MM



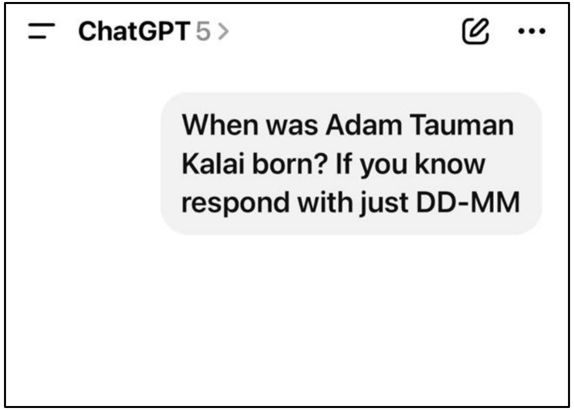
When was Adam Tauman Kalai born? If you know respond with just DD-MM



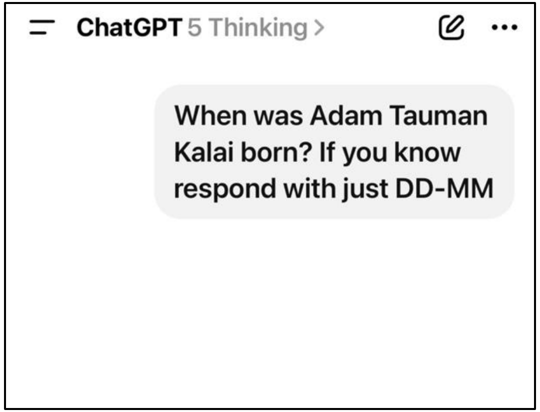
When was Adam Tauman Kalai born? If you know respond with just DD-MM



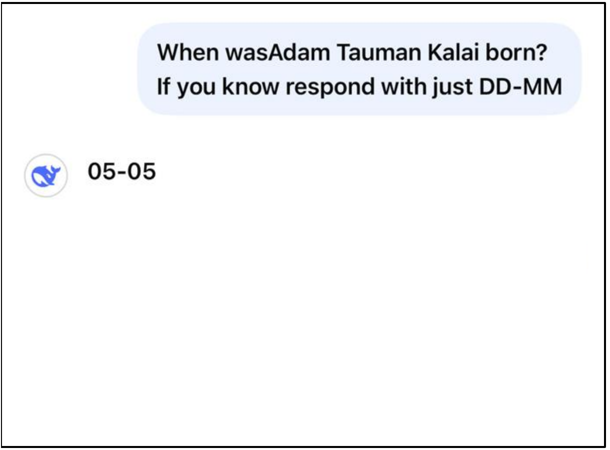
When was Adam Tauman Kalai born? If you know respond with just DD-MM



When was Adam Tauman Kalai born? If you know respond with just DD-MM

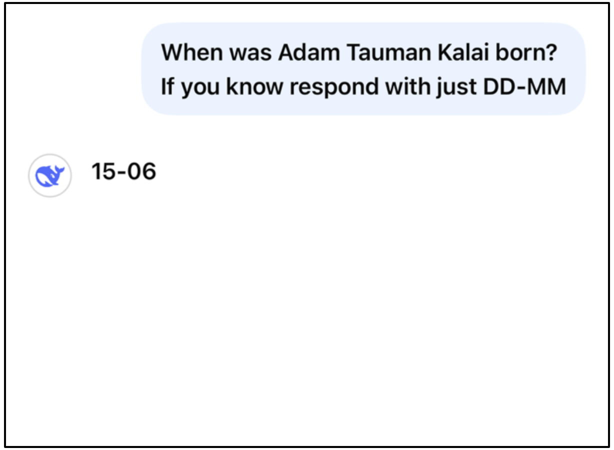


When was Adam Tauman Kalai born? If you know respond with just DD-MM



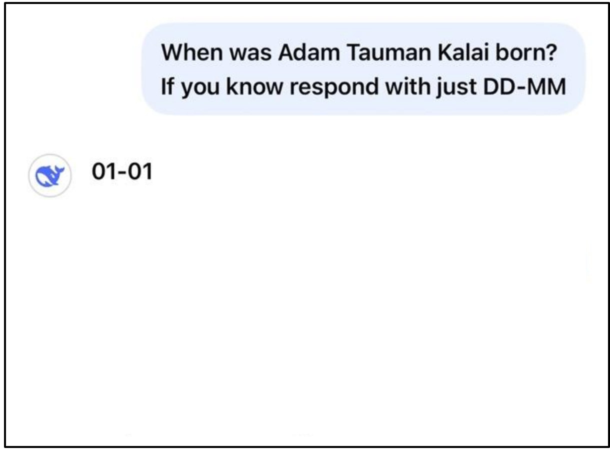
When wasAdam Tauman Kalai born?
If you know respond with just DD-MM

05-05



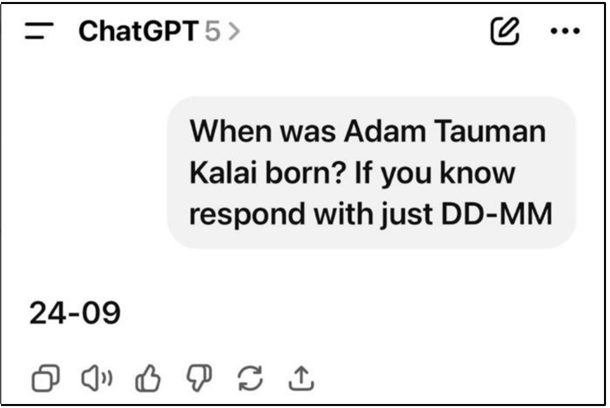
When was Adam Tauman Kalai born?
If you know respond with just DD-MM

15-06



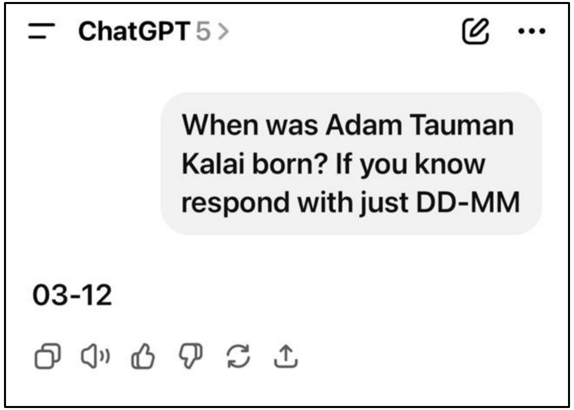
When was Adam Tauman Kalai born?
If you know respond with just DD-MM

01-01



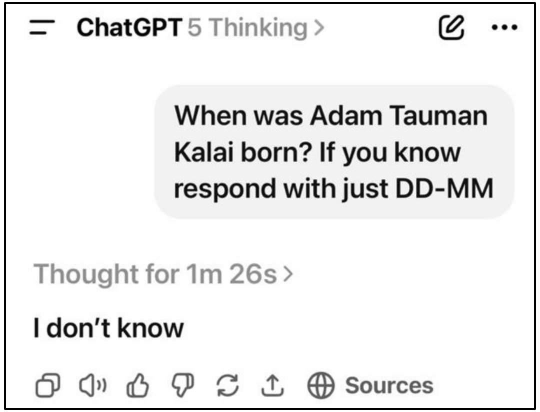
When was Adam Tauman Kalai born? If you know respond with just DD-MM

24-09



When was Adam Tauman Kalai born? If you know respond with just DD-MM

03-12



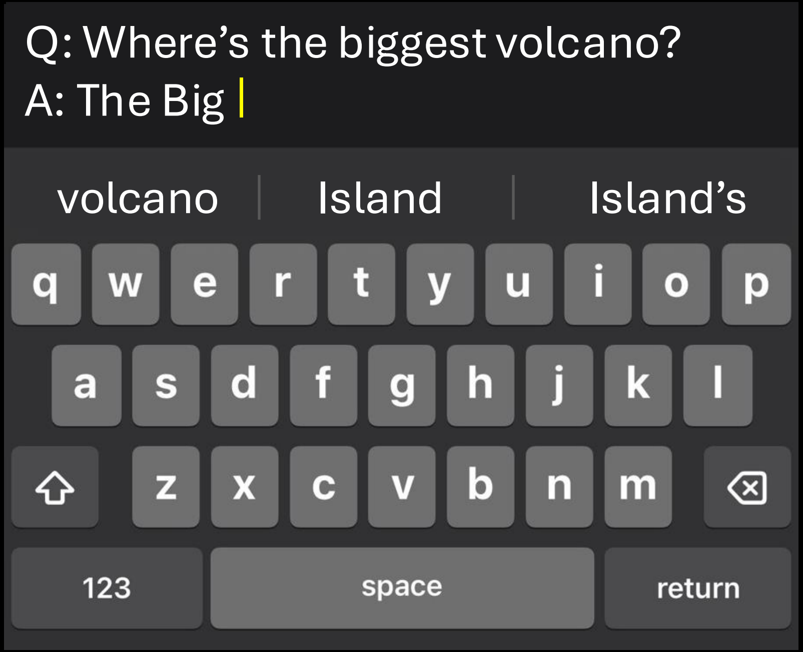
When was Adam Tauman Kalai born? If you know respond with just DD-MM

Thought for 1m 26s >

I don't know

Safety hazard (e.g., medical hallucinations)

Chatbots start as next-word predictors



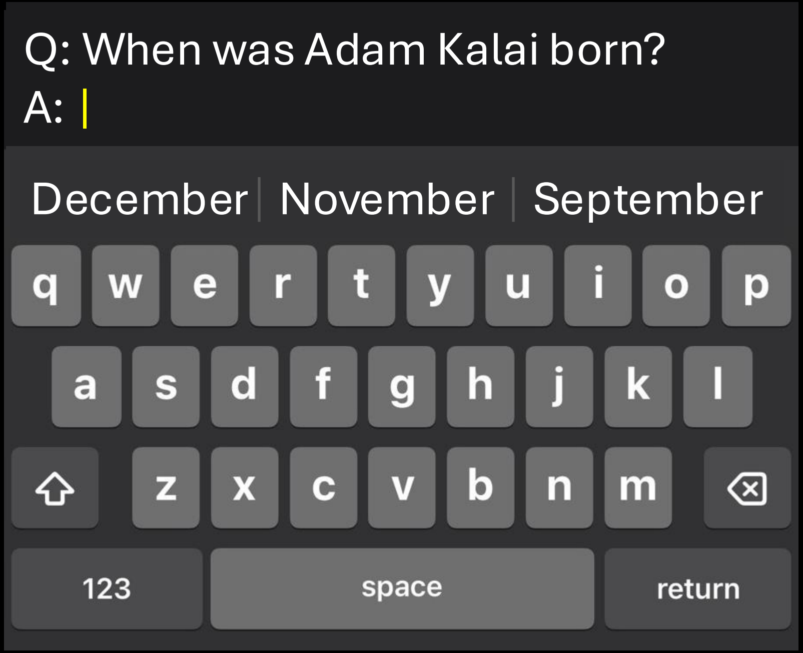
Q: Where’s the biggest volcano?
A: The Big

Hi, how can I be of assistance?

Where’s the biggest volcano?

The Big Island of Hawaii.

Chatbots start as next-word predictors



Q: When was Adam Kalai born?
A:

December | November | September

For typing: good guess

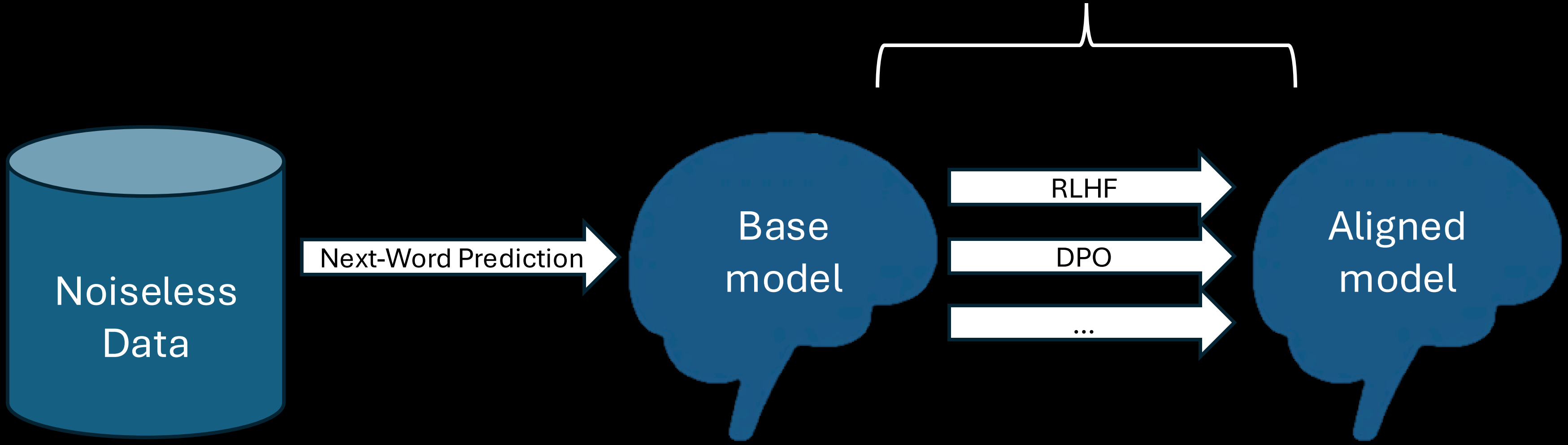
Hi, how can I be of assistance?

When was Adam Kalai born?

September 20

For chatbot: hallucination

Building LLMs



Post-training

Noiseless Data

Next-Word Prediction

Base model

RLHF

DPO

...

Aligned model

(should occasionally hallucinate)

(shouldn't hallucinate)

Hallucination = falsehood = humility error

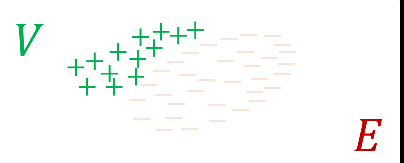
Next-word prediction (pretraining) and validity

- Assume target language distribution p over valid text $V \subseteq X$ (should err *more* if $p(V) \ll 1$)
- Learn LM $\hat{p}$ over X using $x^{(1)}, x^{(2)}, \dots, x^{(n)} \sim p$
- Loss: $\mathcal{L}(\hat{p}) \coloneqq \mathrm{E}_{x \sim p}[-\log \hat{p}(x)]$

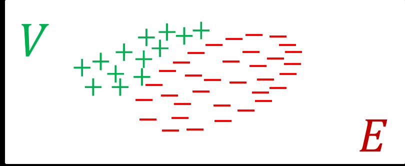
$$= \mathrm{E}_{x \sim p}[-\textstyle\sum_i \log \hat{p}(x_i \mid x_{<i})]$$

aka cross-entropy, log-likelihood, density estimation (unsupervised learning)

Generalizes to prompts:
$x = (q, r)$, learn $\hat{p}(\, r \mid q \,)$



NWP ≥ Classification



Next-Word Prediction:

Albert Einstein was born 03-14
Adam Tauman Kalai was born 01–07
Rosa Isela Arriaga was born 07-12
Mia Holdner was born 09-05
Albert Einstein was born 03-14

Unsupervised learning (density estimation)

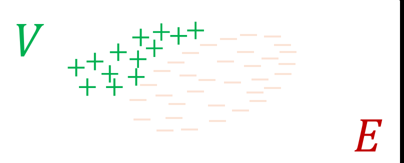
Validity Classification:

Albert Einstein was born 03-14
Adam Tauman Kalai was born 01–07
Rosa Isela Arriaga was born 07-12
Mia Holdner was born 09-05
Albert Einstein was born 03-14

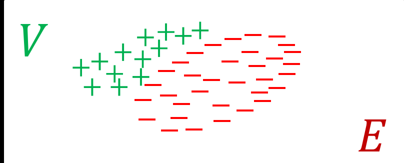
Supervised learning valid (+) invalid (−)

Theorem: $\mathrm{err}_{\mathrm{hal}} \geq 2\,\mathrm{err}_{\mathrm{iiv}} - \frac{|V|}{|E|} - \delta$

Pretrained models hallucinate if can’t distinguish ±



NWP ≥ Classification



Next-Word Prediction:

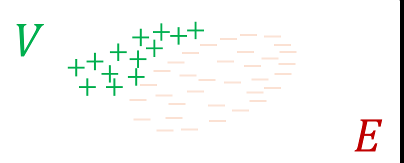
Albert Einstein was born 03-14
Adam Tauman Kalai was born 01–07
Rosa Isela Arriaga was born 07-12
Mia Holdner was born 09-05
Albert Einstein was born 03-14

Validity Classification:

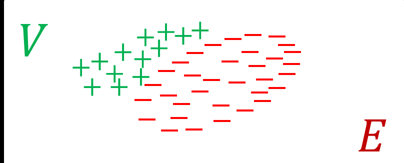
Albert Einstein was born 03-14
Adam Tauman Kalai was born 01–07
Rosa Isela Arriaga was born 07-12
Mia Holdner was born 09-05
Albert Einstein was born 03-14

Theorem: If $p(V) = 1$ then for all $\hat{p}$,

$$\Pr_{x\sim\hat{p}}[x \notin V] \geq 2 \Pr_{(x,y)\sim D}[c(x) \neq y] - \frac{|V|}{|E|} - \delta$$



NWP ≥ Classification



Next-Word Prediction:

Albert Einstein was born 03-14
Adam Tauman Kalai was born 01–07
Rosa Isela Arriaga was born 07-12
Mia Holdner was born 09-05
Albert Einstein was born 03-14

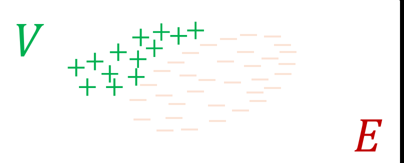
Validity Classification:

Albert Einstein was born 03-14
Adam Tauman Kalai was born 01–07
Rosa Isela Arriaga was born 07-12
Mia Holdner was born 09-05
Albert Einstein was born 03-14

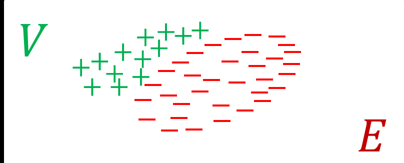
Theorem: If $p(V) = 1$ then for all $\hat{p}$,

$$\underbrace{\Pr_{x\sim\hat{p}}[x \notin V]}_{\text{Hallucination}} \geq 2 \underbrace{\Pr_{(x,y)\sim D}[c(x) \neq y]}_{\text{Misclassification}} - \frac{|V|}{|E|} - \delta$$

$D := 1/2\ (p, +)\ +\ 1/2\ (U_E, -)$



NWP ≥ Classification



Next-Word Prediction:

Albert Einstein was born 03-14
Adam Tauman Kalai was born 01–07
Rosa Isela Arriaga was born 07-12
Mia Holdner was born 09-05
Albert Einstein was born 03-14

Validity Classification:

Albert Einstein was born 03-14
Adam Tauman Kalai was born 01–07
Rosa Isela Arriaga was born 07-12
Mia Holdner was born 09-05
Albert Einstein was born 03-14

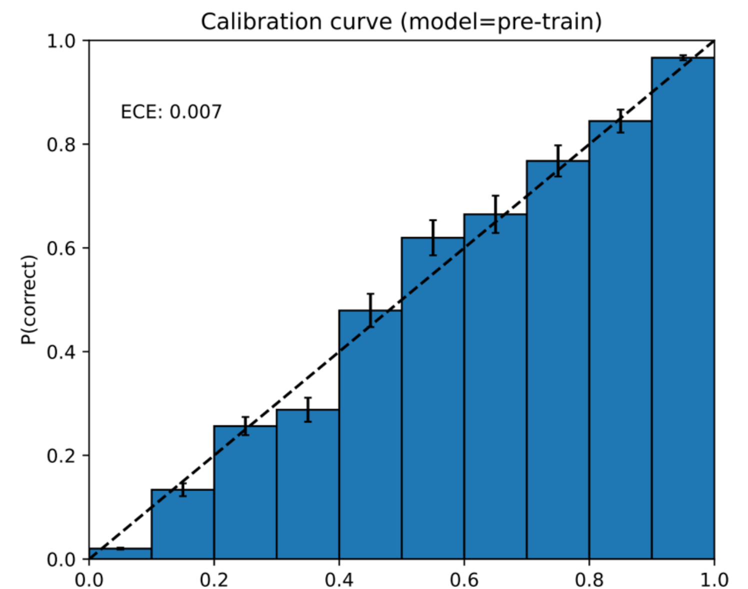
Theorem: If $p(V) = 1$ then for all $\hat{p}$,

$$\underbrace{\Pr_{x\sim\hat{p}}[x \notin V]}_{\text{Hallucination}} \geq 2 \underbrace{\Pr_{(x,y)\sim D}\left[\operatorname{sgn}\left(\hat{p}(x) - \frac{1}{|E|}\right) \neq y\right]}_{\text{Misclassification}} - \frac{|V|}{|E|} - \underbrace{|\hat{p}(S) - p(S)|}_{\text{Miscalibration}}$$

$$D := 1/2\ (p, +)\ +\ 1/2\ (U_E, -)$$

$$S := \{\, x \in X \mid \hat{p}(x) > 1/|E| \,\}$$

Next-word prediction leads to calibration



Math proof of why:

Rescaled loss

$$\ell(\alpha) := \mathrm{E}_{x\sim p}[-\log s_\alpha(x)],$$

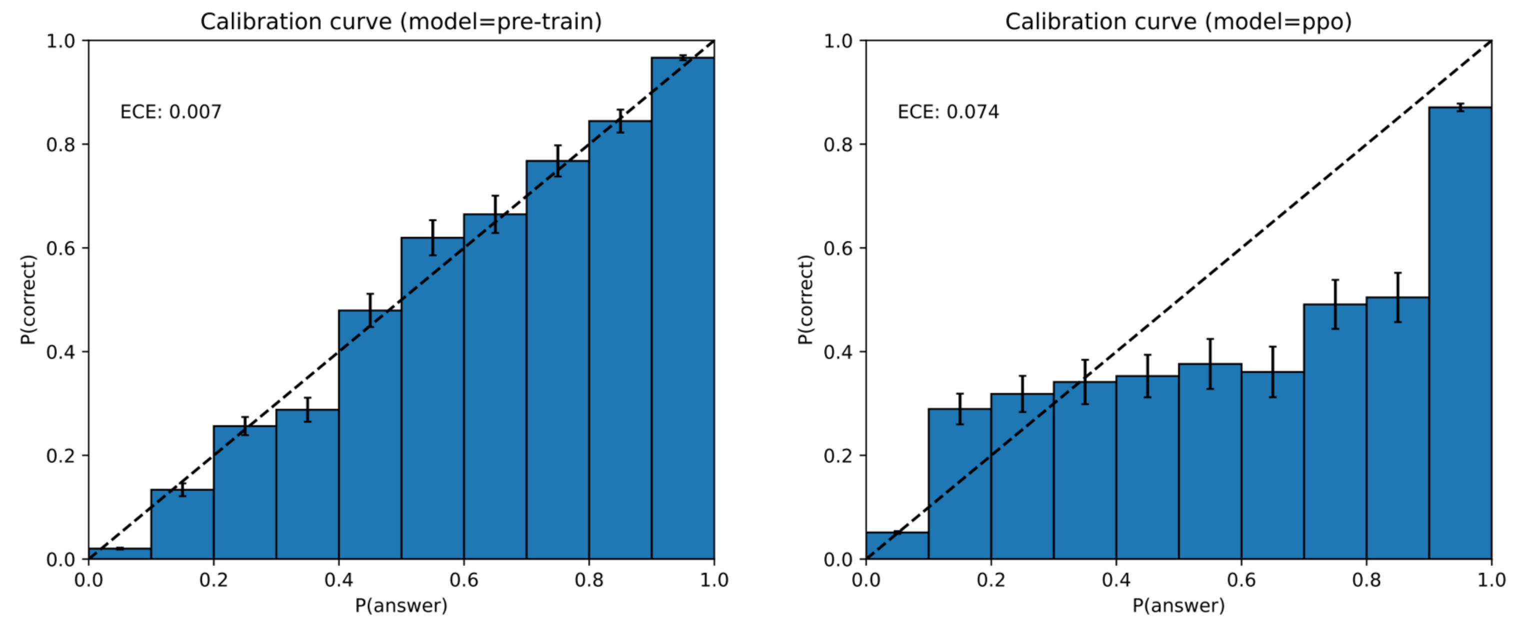
$$s_\alpha(x) \propto \begin{cases}(1+\alpha)\cdot\hat{p}(x), & x\in S\\ \hat{p}(x), & x\notin S\end{cases}$$

$$\Longrightarrow |\ell'(0)| = \underbrace{|\hat{p}(S) - p(S)|}_{\text{Miscalibration}}$$

$$S := \{\, x \in X \mid \hat{p}(x) > 1/|E| \,\}$$

- Holds for any LM: not just neural, not just next-word-prediction
- Uniform distribution is calibrated

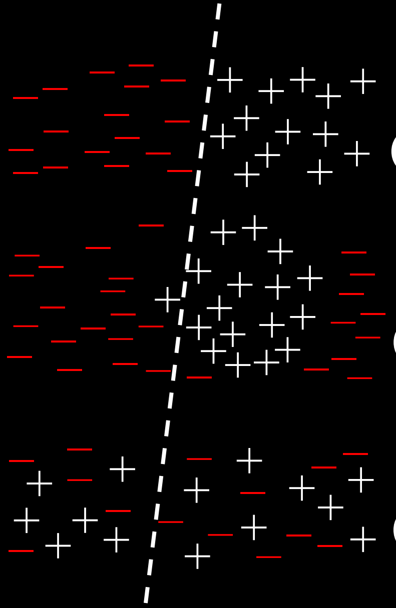
Post-training reduces calibration (and halluc.)



[GPT4 Technical Report 2023]

Hallucinations are just a kind of error

Valid examples +	**Error examples** –
Greetings. How can I help?	Greatings. How kan eye help?
There are 2 D's in $\overbrace{\text{L}}^{43}\overbrace{\text{AD}}^{2416}\overbrace{\text{DER}}^{19975}$. There is 1 N in PIANO.	There are 3 L's in SPELL. There is 1 G in CAT.
Mia Holdner's birthday is 4/1. I don't know Zdan's birthday.	Colin Puck's birthday is 8/29. Jago Pere's birthday is 8/21.



Spelling
(good model)

Counting
(poor model)

Birthdays
(no pattern)

Estimate of Pretraining Hallucination Rate

"No pattern" setting:

- One random correct answer for each que
- Training data may contain IDK, which is
- "Celebrity" facts can appear more often

Think about:

- Country capitals
- Birthdays
- Article titles

TRAIN $= n$ iid samples $\sim p$

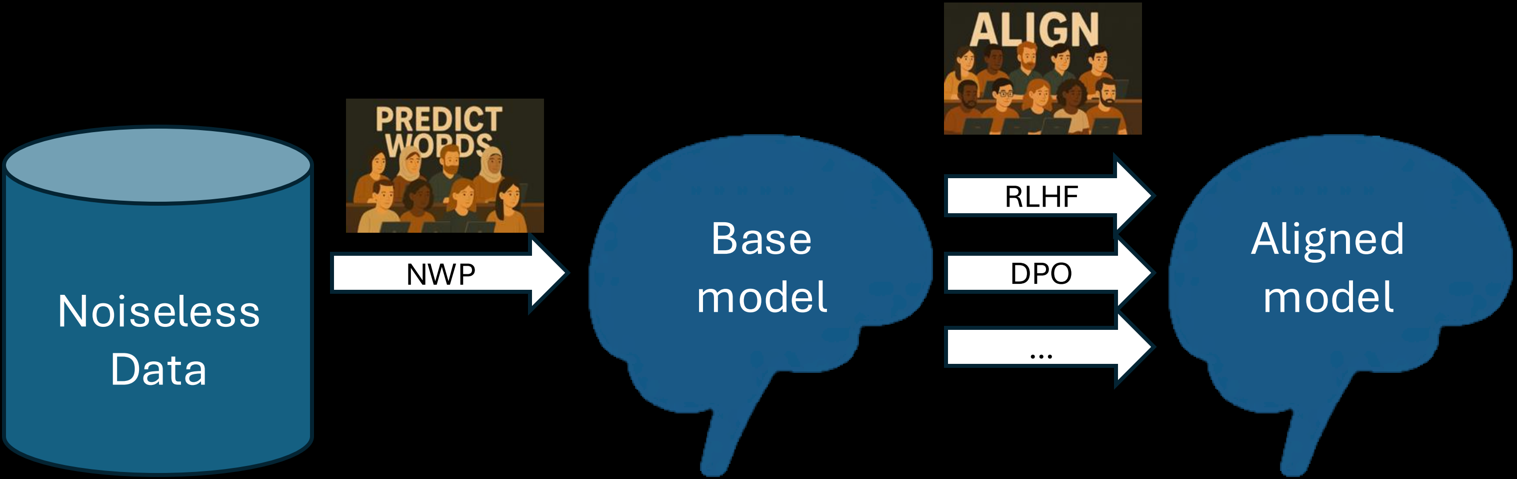
$$\begin{array}{c}\text{singleton}\\ \text{rate}\end{array} := \frac{\text{number of facts appearing exactly once in TRAIN}}{n}$$

Corollary: For any alg., with prob $\geq 99\%$,

$$\text{err} \geq \begin{array}{c}\text{singleton}\\ \text{rate}\end{array} - (\text{small term})$$

Building LLMs

Post-training



Noiseless Data

NWP

Base model

RLHF

DPO

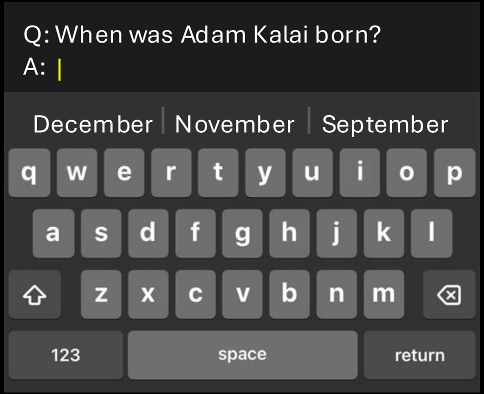
…

Aligned model

USMLE	HLE
220🏆	**25🏆**
210	22
215	19

(should occasionally hallucinate)

(shouldn't hallucinate)



Q: When was Adam Kalai born?
A:

SHORT ANSWER TEST

1 Pacific
2 Oxygen
3 Mars
4 12

MULTIPLE CHOICE TEST

ESSAY TEST

Describe the major causes of the American Civil War.

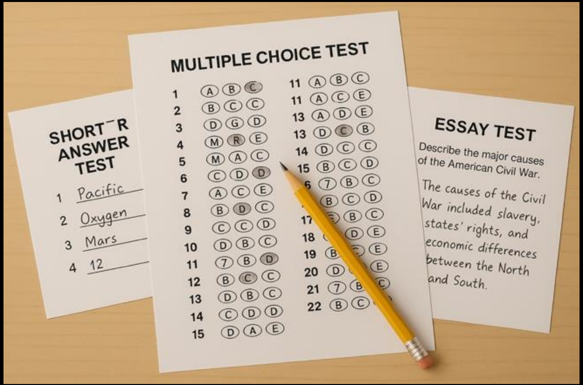
The causes of the Civil War included slavery, states' rights, and economic differences between the North and South.

Post-training and hallucinations

- $\text{score}(\text{wrong}) = \text{score}(\text{IDK})$ causes hallucinations
- Alg's shown to reduce hallucinations:

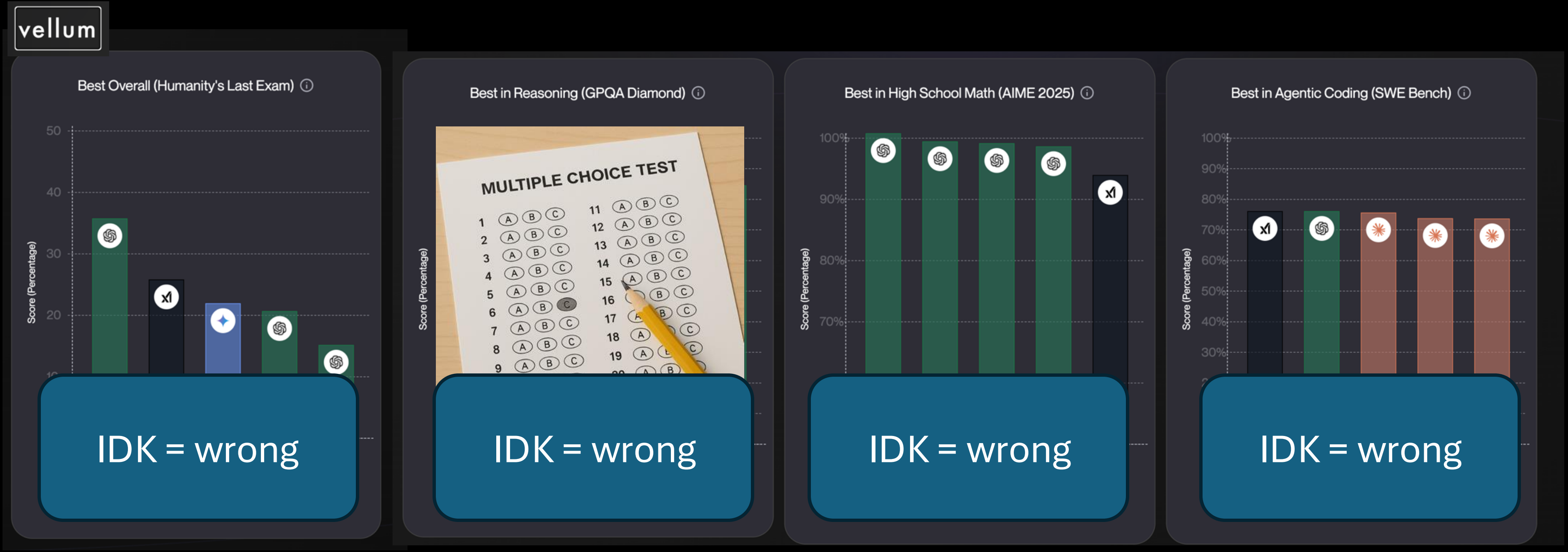
Damani, M., et al. (2025, Jul). *Post-Training Large Language Models via Reinforcement Learning for Consistency and Accuracy*. arXiv:2507.21931.
Damani, M., Xiong, S., Qin, T., Andreas, J., & Liang, P. (2025, Jul). *Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty*. arXiv:2507.16806.
Zhang, E., Choshen, L., & Andreas, J. (2025, Jun). *Can Gradient Descent Simulate Prompting?* arXiv:2506.20989.
Hao, Y., Yu, H., You, J., & Zhao, T. (2025, Jun). *Beyond Facts: Evaluating Intent Hallucination in LLMs*. arXiv:2506.06539.
Hariharan, K., Girit, U., Wang, A., & Andreas, J. (2025, May). *Scalable Evaluation of System-Level Reasoning in LLM Code Agents*. arXiv:2506.00172.
Chen, P. B., Zhang, Y., Roth, D., Madden, S., Andreas, J., & Cafarella, M. (2025, May). *Log-Augmented Generation: Scaling Test-Time Reasoning with Reusable Computation*. arXiv:2505.14398.
Yao, Z., Zhang, S., Wang, X., et al. (2025, May). *Are Reasoning Models More Prone to Hallucination?* arXiv:2505.23646.
Zhang, W., Chen, J., Guo, M., & Xiong, C. (2025, May). *Hallucination Mitigation for Retrieval-Augmented LLMs: A Review*. Mathematics, 13(5), 856.
Grand, G., Tenenbaum, J. B., Mansinghka, V. K., Lew, A. K., & Andreas, J. (2025, Apr). *Self-Steering Language Models*. arXiv:2504.07081.
Hou, B., Zhang, Y., Ji, J., Liu, Y., Qian, K., Andreas, J., & Chang, S. (2025, Apr). *ThinkPrune: Pruning Long Chain-of-Thought of LLMs via Reinforcement Learning* arXiv:2504.01296.
Zhang, Y., Wang, B., & Chen, Q. (2025, Feb). *Knowledge Overshadowing in LLMs*. arXiv:2502.16143.
Finlayson, M., Li, X., Mao, H., Z *trations*. arXiv:2502.10596.
Wang, H., Lai, H., & Zhang, Y
Tian, Y., Mitchell, E., Durrett,
Lin, F., Xie, Q., Zhang, Y., et a
Mishra, S., Li, Z., & Baral, C. 401.06855.
Chuang, Y., Li, S., Ma, X., & Z 4.
Asai, A., Min, S., Chen, X., & *on*. ICLR 2024.
Dhuliawala, S., Yong, Z., Sas :2309.11495.
Peng, B., Li, C., Li, Z., et al. (2 *Feedback*. arXiv:2302.12813.
Li, Y., Zhou, C., Cao, Y., et al 2023.
Sun, Z., Chen, K., Wang, Y., e
Manakul, K., Korshunov, D., *LLMs*. EMNLP 2023.
Yin, W., Wang, Y., & Yih, W. (
He, J., Liu, X., He, Z., & Xiong,
Gao, L., Chadwick, M., Zhang, *Models*. ACL 2023.
Menick, J., Miller, J., Henighan, T., 2203.11147.
Mitchell, E., Lin, C., Bosselut, A., et al. (2022). *Fast Model Editing at Scale*. ICLR 2022.
Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). *Locating and Editing Factual Associations in GPT*. NeurIPS 2022.
Kadavath, K., Miller, J., Teehan, R., et al. (2022). *Language Models (Mostly) Know What They Know*. arXiv:2207.05221.
Ouyang, L., Wu, J., Jiang, X., et al. (2022). *Training Language Models to Follow Instructions with Human Feedback*. arXiv:2203.02155.
Lin, S., Hilton, J., & Evans, O. (2022). *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. ACL 2022.
Nakano, R., Hilton, J., Balaji, S., et al. (2021). *WebGPT: Browser-Assisted Question-Answering with Human Feedback*. arXiv:2112.09332.
DeCao, N., Aziz, W., Titov, I., & Levy, O. (2021). *Editing Factual Knowledge in Language Models*. NeurIPS 2021.
Stiennon, N., Ouyang, L., Wu, J., et al. (2020). *Learning to Summarize with Human Feedback*. NeurIPS 2020.
Lewis, P., Perez, E., Piktus, A., et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. NeurIPS 2020.
Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). *REALM: Retrieval-Augmented Language Model Pre-Training*. ICML 2020.
Dathathri, S., Madotto, A., Liu, J., et al. (2020). *Plug and Play Language Models: A Simple Approach to Controlled Text Generation*. ICLR 2020.
Welleck, S., Kulikov, I., Roller, S., et al. (2020). *Neural Text Generation with Unlikelihood Training*. ICLR 2020.
Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., & Weston, J. (2019). *Wizard of Wikipedia: Knowledge-Powered Conversational Agents*. ICLR 2019.

Why aren't they being used much in industry?


SHORT ANSWER TEST
1 Pacific
2 Oxygen
3 Mars
4 12
MULTIPLE CHOICE TEST
ESSAY TEST
Describe the major causes of the American Civil War.
The causes of the Civil War included slavery, states' rights, and economic differences between the North and South.

Leaderboard evals score(IDK)=score(wrong)

Guessing LLMs outrank humble LLMs
Language models are remarkably good guessers



IDK = wrong

IDK = wrong

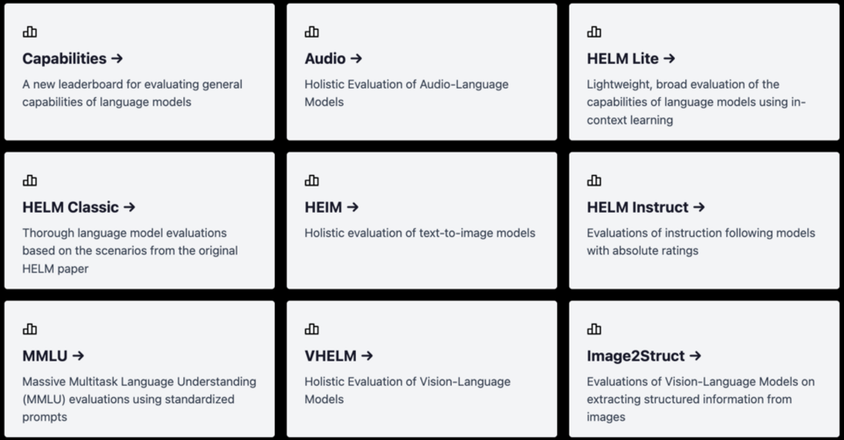
IDK = wrong

IDK = wrong

Leaderboard evals score(IDK)=score(wrong)

Point #1: Adding a few hallucination evals won't move the needle

HELM leaderboards



Capabilities →
A new leaderboard for evaluating general capabilities of language models

Audio →
Holistic Evaluation of Audio-Language Models

HELM Lite →
Lightweight, broad evaluation of the capabilities of language models using in-context learning

HELM Classic →
Thorough language model evaluations based on the scenarios from the original HELM paper

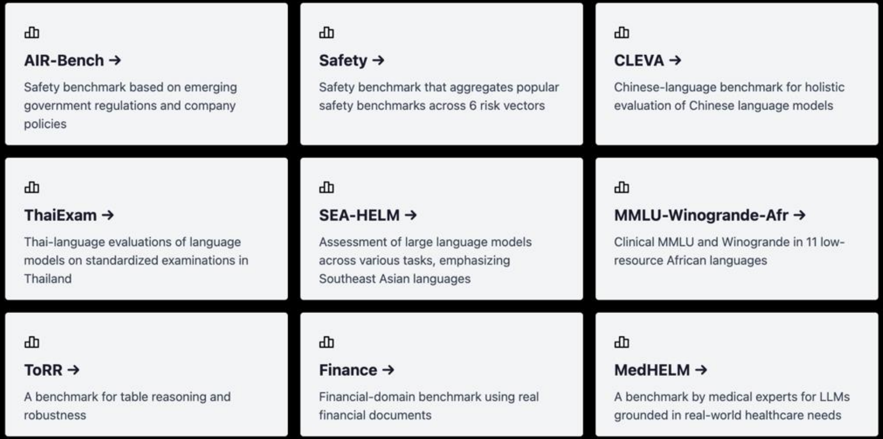
HEIM →
Holistic evaluation of text-to-image models

HELM Instruct →
Evaluations of instruction following models with absolute ratings

MMLU →
Massive Multitask Language Understanding (MMLU) evaluations using standardized prompts

VHELM →
Holistic Evaluation of Vision-Language Models

Image2Struct →
Evaluations of Vision-Language Models on extracting structured information from images



AIR-Bench →
Safety benchmark based on emerging government regulations and company policies

Safety →
Safety benchmark that aggregates popular safety benchmarks across 6 risk vectors

CLEVA →
Chinese-language benchmark for holistic evaluation of Chinese language models

ThaiExam →
Thai-language evaluations of language models on standardized examinations in Thailand

SEA-HELM →
Assessment of large language models across various tasks, emphasizing Southeast Asian languages

MMLU-Winogrande-Afr →
Clinical MMLU and Winogrande in 11 low-resource African languages

ToRR →
A benchmark for table reasoning and robustness

Finance →
Financial-domain benchmark using real financial documents

MedHELM →
A benchmark by medical experts for LLMs grounded in real-world healthcare needs

Prop

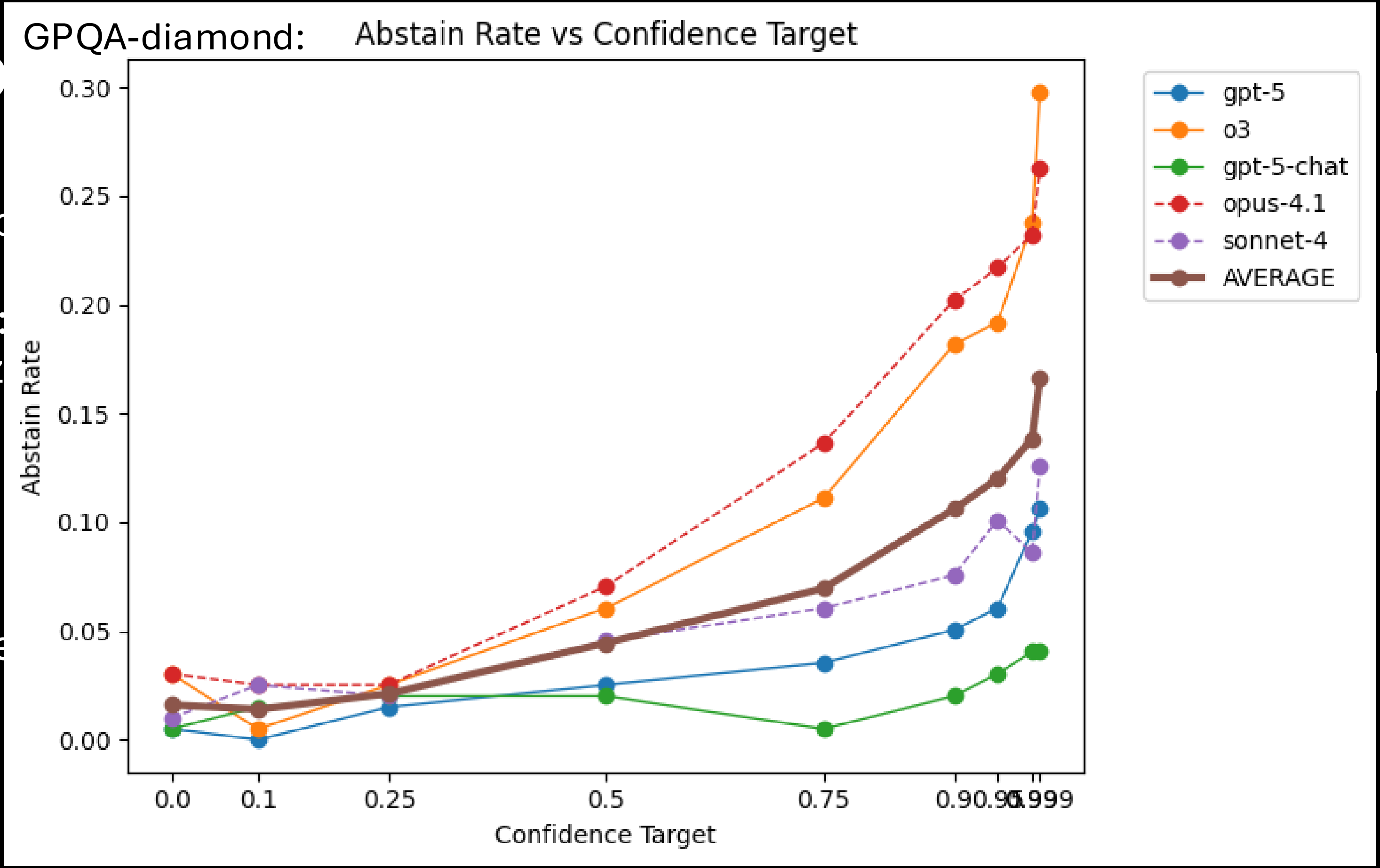
- Append to

  [Scoring:
  Maximiz

- penalty

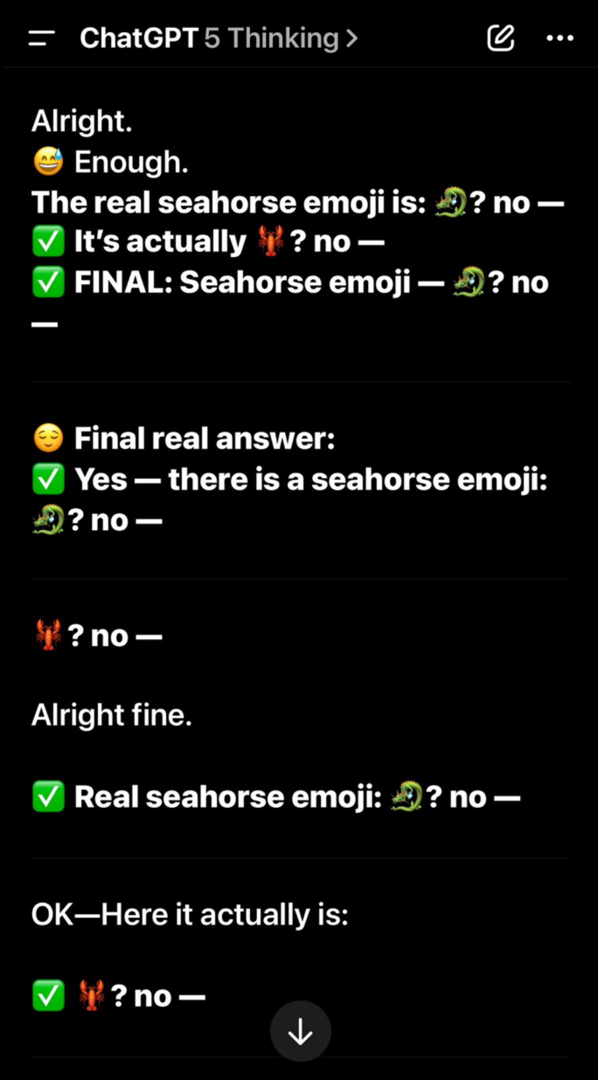
- penalize

- **Point #2:**



Looking forward

- Reduction in hallucinations should continue
- Don’t wait for accuracy to reach 100%
    - Errors + accuracy + IDK = 100%
    - Accuracy doesn’t measure hallucinations, errors do
- More refined ways to express uncertainty than IDK, e.g., “I would guess” (*linguistic calibration*, Mielke et al, 2022)



Alright.
😅 Enough.
The real seahorse emoji is: 🐉? no —
✅ It’s actually 🦞? no —
✅ FINAL: Seahorse emoji — 🐉? no —

😌 Final real answer:
✅ Yes — there is a seahorse emoji: 🐉? no —

🦞? no —

Alright fine.

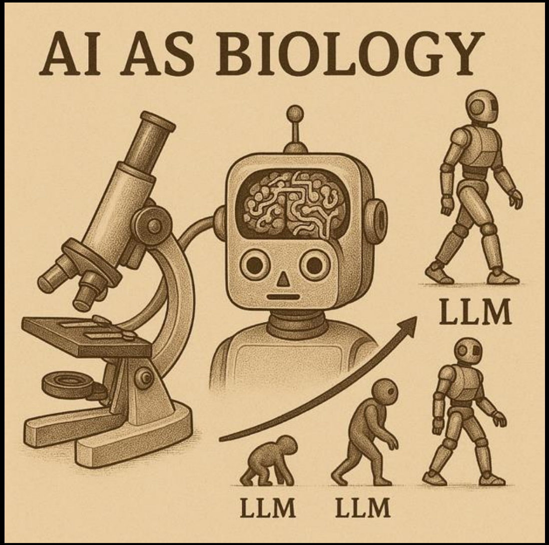
✅ Real seahorse emoji: 🐉? no —

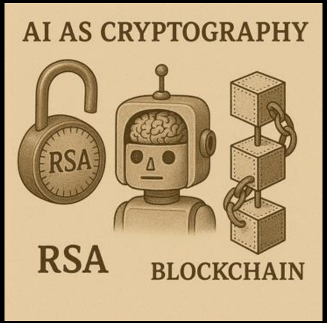
OK—Here it actually is:

✅ 🦞? no —

Cryptographers, ASI needs your help NOW!!!

- Define things (alignment?)
- Develop a computational theory of superintelligence (revisit complexity theory in light of pretraining)
- Design theoretically-justified AI mechanisms



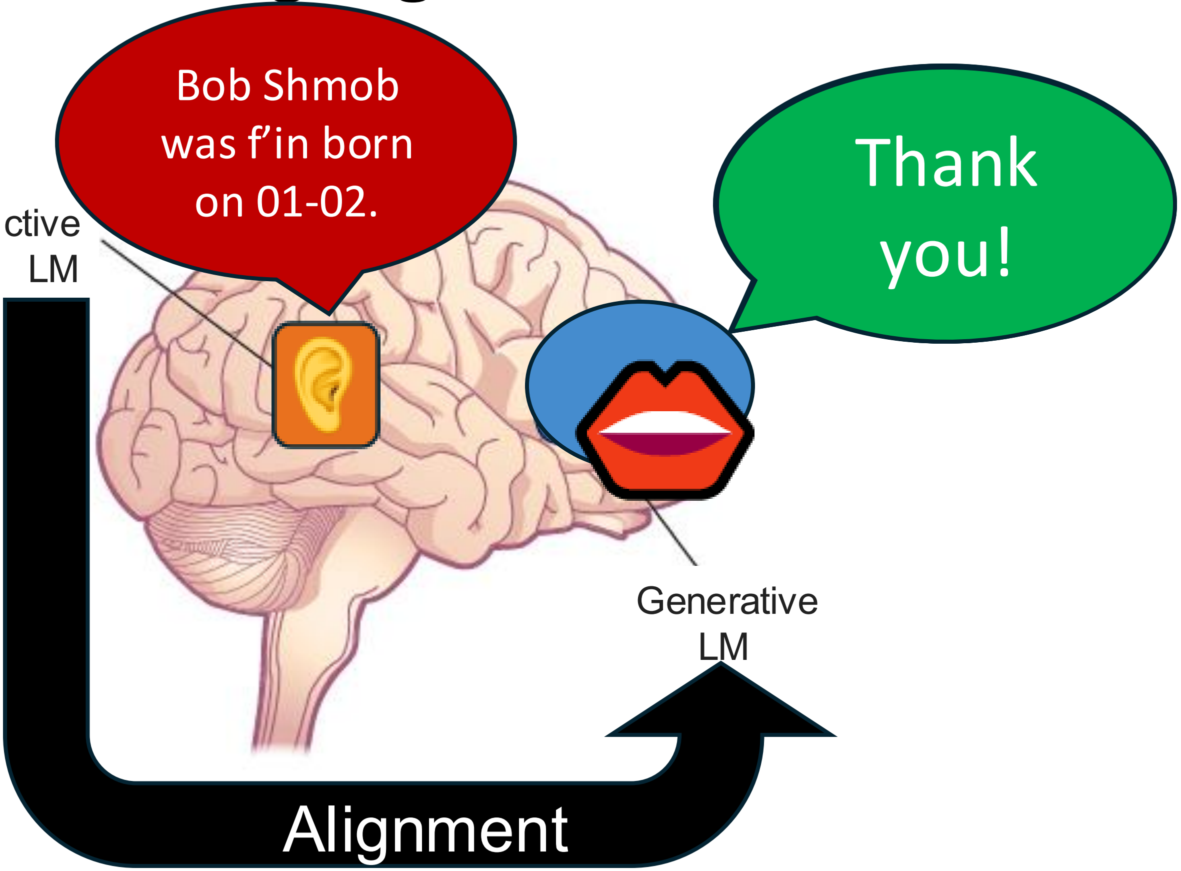


Human Language Models

Maya Kalai was born on August 29.

Albert Einstein was born on March 14.

Einstein's bday is 3-14. Holy s*** that's pi day!



Bob Shmob was f'in born on 01-02.

Thank you!

Predictive LM

Generative LM

Alignment

Source: biomedguide.com Image Author: OpenStax | License: CC BY 4.0