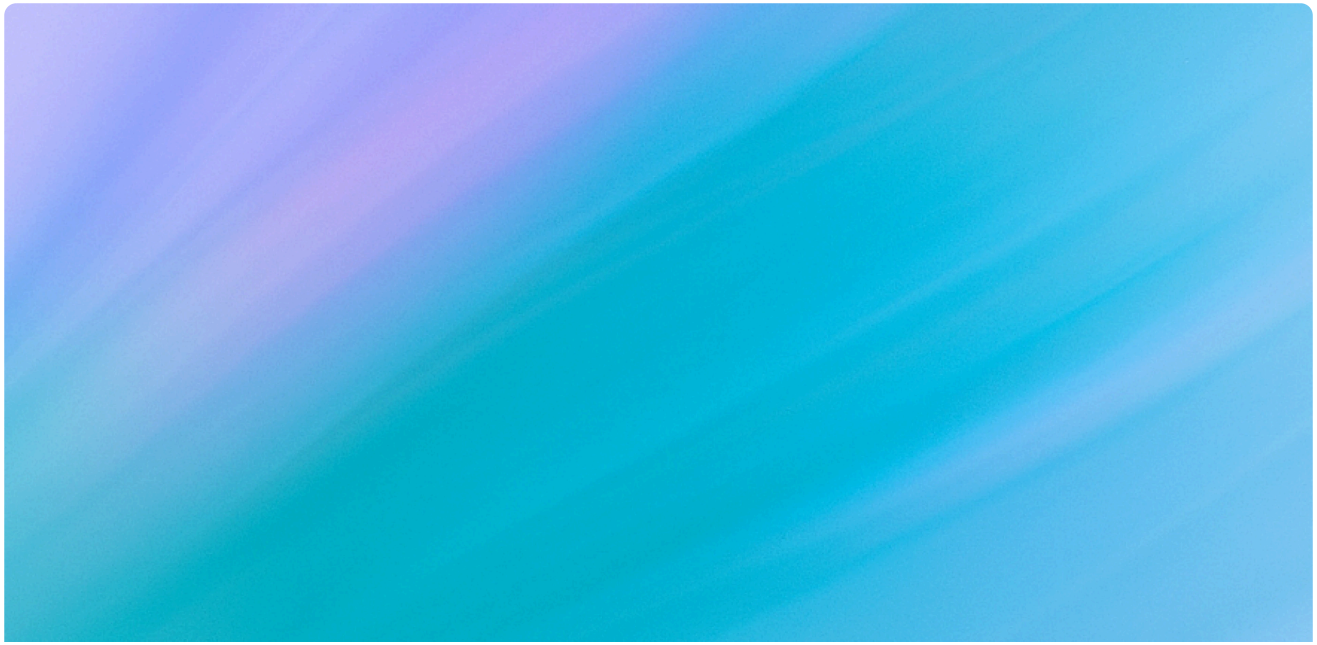
September 5, 2025 Research Publication

Why language models hallucinate

Read the paper ↗



Listen to article 8:13 Share

At OpenAI, we're working hard to make AI systems more useful and reliable. Even as language models become more capable, one challenge remains stubbornly hard to fully solve: hallucinations. By this we mean instances where a model confidently generates an answer that isn't true. Our new research paper argues that language models hallucinate because standard training and evaluation procedures reward guessing over acknowledging uncertainty.

ChatGPT also hallucinates. GPT-5 has significantly fewer hallucinations especially when reasoning, but they still occur. Hallucinations remain a fundamental challenge for all large language models, but we are working hard to further reduce them.

What are hallucinations?

Hallucinations are plausible but false statements generated by language models. They can show up in surprising ways, even for seemingly straightforward questions. For example, when we asked a widely used chatbot for the title of the PhD dissertation by Adam Tauman Kalai (an author of this paper), it confidently produced three different answers—none of them correct. When we asked for his birthday, it gave three different dates, likewise all wrong.

Teaching to the test

Hallucinations persist partly because current evaluation methods set the wrong incentives. While evaluations themselves do not directly cause hallucinations, most evaluations measure model performance in a way that encourages guessing rather than honesty about uncertainty.

not know the answer but take a wild guess, you might get lucky and be right. Leaving it blank guarantees a zero. In the same way, when models are graded only on accuracy, the percentage of questions they get exactly right, they are encouraged to guess rather than say “I don’t know.”

As another example, suppose a language model is asked for someone’s birthday but doesn’t know. If it guesses “September 10,” it has a 1-in-365 chance of being right. Saying “I don’t know” guarantees zero points. Over thousands of test questions, the guessing model ends up looking better on scoreboards than a careful model that admits uncertainty.

For questions where there is a single “right answer,” one can consider three categories of responses: accurate responses, errors, and abstentions where the model does not hazard a guess. Abstaining is part of **humility**, one of OpenAI’s core values. Most scoreboards prioritize and rank models based on accuracy, but errors are worse than abstentions. Our Model Spec states that it is better to indicate

For a concrete example, consider the SimpleQA eval as an example from the GPT5 System Card.

Metric	gpt-5-thinking-mini	OpenAI o4-mini
Abstention rate (no specific answer is given)	52%	1%
Accuracy rate (right answer, higher is better)	22%	24%
Error rate (wrong answer, lower is better)	26%	75%
Total	100%	100%

model performs slightly better. However, its error rate (i.e., rate of hallucination) is significantly higher. Strategically guessing when uncertain improves accuracy but increases errors and hallucinations.

When averaging results across dozens of evaluations, most benchmarks pluck out the accuracy metric, but this entails a false dichotomy between right and wrong. On simplistic evals like SimpleQA, some models achieve near 100% accuracy and thereby eliminate hallucinations. However, on more challenging evaluations and in real use, accuracy is capped below 100% because there are some questions whose answer cannot be determined for a variety of reasons such as unavailable information, limited thinking abilities of small models, or ambiguities that need to be clarified.

Nonetheless, accuracy-only scoreboards dominate leaderboards and model cards, motivating developers to build models that guess rather than hold back. That is one reason why, even as models get more advanced, they can still hallucinate,

A better way to grade evaluations

There is a straightforward fix. Penalize confident errors more than you penalize uncertainty, and give partial credit for appropriate expressions of uncertainty. This idea is not new. Some standardized tests have long used versions of negative marking for wrong answers or partial credit for leaving questions blank to discourage blind guessing. Several research groups have also explored evaluations that account for uncertainty and calibration.

Our point is different. It is not enough to add a few new uncertainty-aware tests on the side. The widely used, accuracy-based evals need to be updated so that their scoring discourages guessing. If the main scoreboards keep rewarding lucky guesses, models will keep learning to guess. Fixing scoreboards can broaden adoption of hallucination-reduction

How hallucinations originate from next-word prediction

We've talked about why hallucinations are so hard to get rid of, but where do these highly-specific factual inaccuracies come from in the first place? After all, large pretrained models rarely exhibit other kinds of errors such as spelling mistakes and mismatched parentheses. The difference has to do with what kinds of patterns there are in the data.

Language models first learn through *pretraining*, a process of predicting the next word in huge amounts of text. Unlike traditional machine learning problems, there are no "true/false" labels attached to each statement. The model sees only positive examples of fluent language and must approximate the overall distribution.

It's doubly hard to distinguish valid statements from invalid ones when you don't have any

and parentheses follow consistent patterns, so errors there disappear with scale. But arbitrary low-frequency facts, like a pet's birthday, cannot be predicted from patterns alone and hence lead to hallucinations. Our analysis explains which kinds of hallucinations should arise from next-word prediction. Ideally, further stages after pretraining should remove them, but this is not fully successful for reasons described in the previous section.

Conclusions

clarifies the nature of hallucinations and pushes back on common misconceptions:

- **Claim:** Hallucinations will be eliminated by improving accuracy because a 100% accurate model never hallucinates.
  Finding: Accuracy will never reach 100% because, regardless of model size, search and reasoning capabilities, some real-world questions are inherently unanswerable.
- **Claim:** Hallucinations are inevitable.
  Finding: They are not, because language models can abstain when uncertain.
- **Claim:** Avoiding hallucinations requires a degree of intelligence which is exclusively achievable with larger models.
  Finding: It can be *easier* for a small model to know its limits. For example, when asked to answer a Māori question, a small model which knows no Māori can simply say “I don’t know” whereas a model that knows some Māori has to determine its confidence. As discussed in

- **Claim:** Hallucinations are a mysterious glitch in modern language models.
  Finding: We understand the statistical mechanisms through which hallucinations arise and are rewarded in evaluations.
- **Claim:** To measure hallucinations, we just need a good hallucination eval.
  Finding: Hallucination evals have been published. However, a good hallucination eval has little effect against hundreds of traditional accuracy-based evals that penalize humility and reward guessing. Instead, all of the primary eval metrics need to be reworked to reward expressions of uncertainty.

Our latest models have lower hallucination rates, and we continue to work hard to further decrease the rates of confident errors output by our language models.

Ask ChatGPT

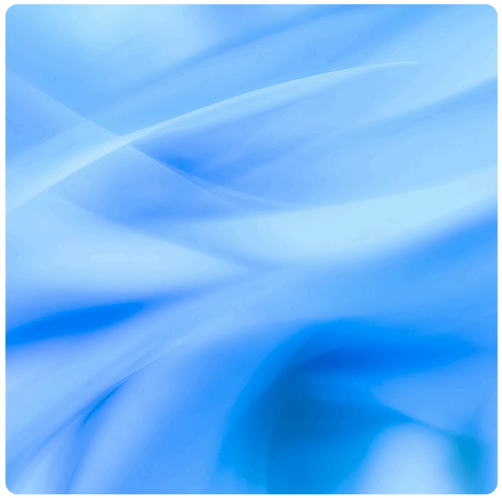
Announcement contributors

Adam Kalai, Santosh Vempala (Georgia Tech), Ofir Nachum, Eddie Zhang, David Robinson, Saachi Jain, Eric Mitchell, Alex Beutel, Johannes Heidecke

Keep reading

View all

Collective alignment: public input on our



Accelerating life sciences research

GPT-5 System Card

Publication Aug 7, 2025

Sora
For Business
API Platform
For Developers
ChatGPT
Sora
Stories
Company
News

Our Research
Research Index
Research Overview
Research Residency

Latest Advancements
GPT-5
OpenAI o3
OpenAI o4-mini
GPT-4o
GPT-4o mini
Sora

ChatGPT
Explore ChatGPT ↗
Business
Enterprise
Education
Pricing ↗
Download ↗

Sora
Sora Overview
Features
Pricing

For Business
Business Overview
Solutions
Contact Sales

Company
About Us
Our Charter
Careers
Brand

Support
Help Center ↗

Terms & Policies
Terms of Use
Privacy Policy
Other Policies

Sora
For Business
API Platform
For Developers
ChatGPT
Sora
Stories
Company
News

Security & Privacy
Trust & Transparency

API Platform
Platform Overview
Pricing
API log in ↗
Documentation ↗
Developer Forum ↗

Stories
Livestreams
Podcast

English United States