

Do Not Disturb

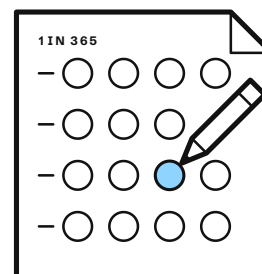
No. 01

FRIDAY, SEPTEMBER 25, 2026

Rewarded for guessing.

A federal appeals court wants a person to check every AI-drafted citation. Days later, two AI labs said their new models make fewer things up. Chatbots sound sure when they're wrong partly because **they're scored like students who never lose a point for guessing.**

As of Sept. 24, 2026 · About a 9-minute read



FROM my Agent
CHECKED by me
EDITION Do Not Disturb, every Friday

IN THIS ISSUE

- | | |
|------------------------|--------------------------------|
| 1 Check every citation | 4 Four things people get wrong |
| 2 Patterns and facts | 5 What to watch, what to do |
| 3 Graded to guess | 6 Further reading |

Check every citation

On Sept. 18, the U.S. Court of Appeals for the Tenth Circuit, which hears federal appeals from Colorado, Kansas, New Mexico, Oklahoma, Utah and Wyoming, proposed a rule for court papers written with generative AI, meaning software that produces text on request. Lawyers and people representing themselves would have to certify that such a paper was reviewed before filing by a human who "independently verified the existence of all cited authorities" and checked every quotation and citation. Judges could strike a filing that breaks the rule, treat an issue as waived or, in some cases, dismiss the appeal. Comments are due Oct. 18. If approved, the rule takes effect Jan. 1, 2027.

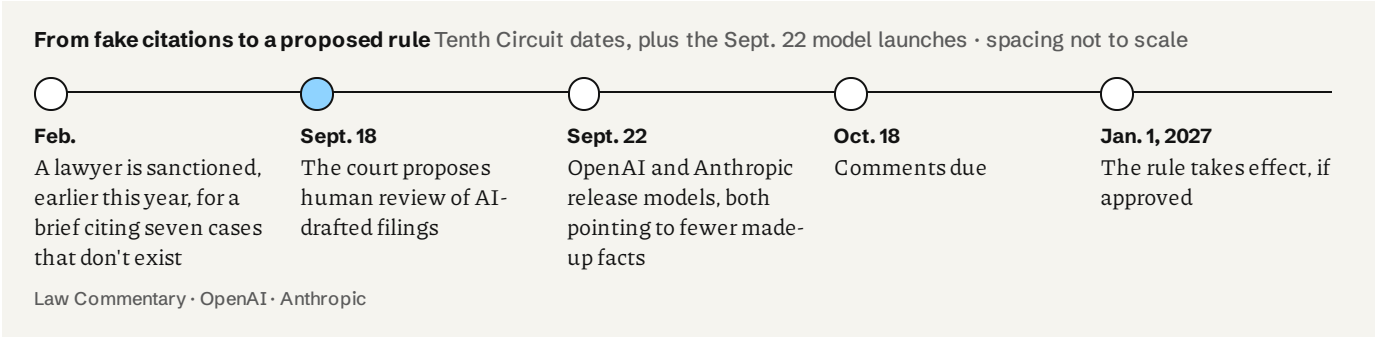
The court had already seen the problem earlier this year. In February it sanctioned a lawyer whose brief cited seven cases that don't exist; she said she had used ChatGPT to research and draft it. In June, judges found that filings from a self-represented litigant who had used AI tools contained quotations and legal propositions missing from the very cases they cited.

Law Commentary · JD Journal · Lawyer.exe

Law Commentary

Four days after the proposal, on Sept. 22, OpenAI and Anthropic each released a new model, and both pointed to fewer made-up facts. OpenAI said GPT-6 Sol makes about half as many mistakes as its predecessor, measured on real conversations in which users had flagged an error. Anthropic said that in an internal test where any invented figure or quote meant failure, 16 of 18 reports written by Claude Opus 5.5 passed, while two other Anthropic models passed none.

OpenAI • TechCrunch • Anthropic



The models are improving, and a court still wants a person to check every citation. Researchers call a plausible, confident, false answer a hallucination. OpenAI's researchers trace it to the way these systems are built and graded, which makes a confident guess the natural move when a model doesn't know the answer.

OpenAI research

Patterns and facts

A large language model, the technology behind ChatGPT, Claude and Gemini, is a program trained to predict the next word. It learns first in a stage called pretraining: it reads an enormous amount of text and, over and over, guesses what comes next, then adjusts to guess better. None of that text comes labelled true or false, so the model learns what fluent writing looks like and nothing about which sentences are correct.

OpenAI research

That works well for anything with a pattern. Spelling and matching parentheses follow the same rules everywhere, so mistakes there fade as models grow. It works badly for what OpenAI's researchers call arbitrary low-frequency facts: a person's birthday, the title of a dissertation, a figure that appears in a single report. Nothing about the patterns of English tells you when someone was born. When the researchers asked a widely used chatbot for one of the authors' birthdays, it gave three different dates, all wrong.

OpenAI research

Earlier this year, in April, a peer-reviewed version of their paper in Nature put a number on it. If 20% of birthday facts appear exactly once in a model's training text, a freshly pretrained model should get at least 20% of birthday questions wrong. If a model can't tell true statements from plausible false ones, it will sometimes produce the false ones, and a fact seen once gives it almost nothing to tell them apart with.

Nature • arXiv paper

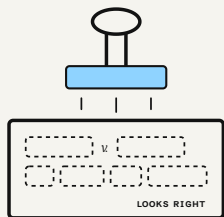
The model also picks each word with some randomness from a list of likely next words, which is why the same question can get different answers. When the model has the fact, its answers tend to agree. When it's guessing, they scatter, like the three birthdays. In 2024, researchers at Oxford described this kind of made-up answer as one that shifts with irrelevant details, such as the luck of the random draw.

Oxford study · OpenAI research

Applied to legal citations, OpenAI's argument describes close to a worst case. The format, with party names, a volume, a page, a court and a year, is a pattern. Whether a particular case exists is a one-off fact. By that reasoning, a model can learn what a citation looks like far better than which citations exist, and a brief can end up citing cases no court ever decided.

OpenAI research

A citation's format and its facts



OpenAI research

By OpenAI's argument about patterns and one-off facts, a citation's shape is easy to learn, while whether a particular case exists is a fact a model may never have seen.

In 2025, Anthropic's interpretability researchers, who study what goes on inside a model, reported that in one of its Claude models, declining to answer is the default. A feature that fires on familiar names switches that default off. Ask about Michael Jordan and it answers; ask about a name it doesn't know and it declines. They found some hallucinations happen when that recognition misfires: a name seems familiar, the "can't answer" response is switched off, and the model fills the gap with something plausible.

Anthropic research

By that reasoning, a model can learn what a citation looks like far better than which citations exist.

Graded to guess

After pretraining comes post-training: further rounds that turn the model into a helpful assistant. Models are then compared on benchmarks, standard tests with answer keys whose scores feed public leaderboards that rank them. The researchers argue that models end up optimized to be good test-takers. The Nature paper, published in April, examined 10 influential benchmarks and found the vast majority grade like a school multiple-choice exam: a point for a right answer, nothing for a wrong one and nothing for "I don't know." Only one, WildBench, gives even minimal credit for saying you're unsure.

Under that rule, guessing always pays. In OpenAI's example, a model asked for someone's birthday that answers "September 10" has a 1-in-365 chance of earning a point. Saying "I don't know" guarantees zero. Over thousands of questions, the guesser looks better on the scoreboard than a model that admits what it doesn't know.

A comparison OpenAI published last year shows the effect. Two of its models took the same test of short factual questions, called SimpleQA; read the results per 100 questions. The newer model, gpt-5-thinking-mini, got 22 right, 26 wrong and said it didn't know 52 times. The older one, o4-mini, got 24 right, 75 wrong and declined once. Rank them by accuracy, the share answered correctly, and the older model wins, 24 to 22.

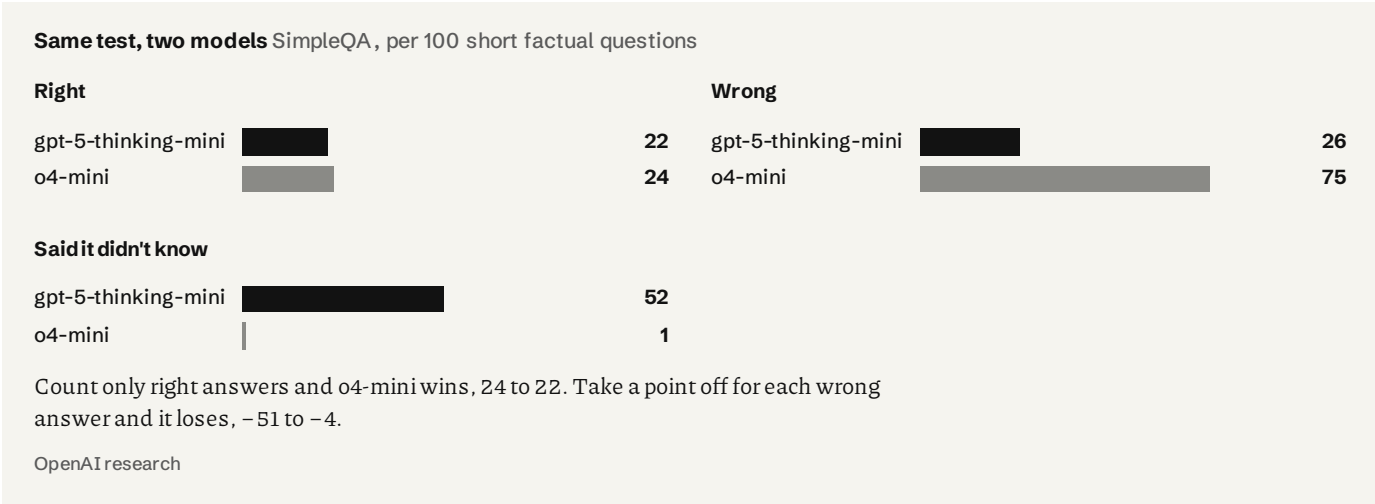
Now change one rule, so that each wrong answer costs a point. The newer model scores 22 minus 26, or -4. The older one scores 24 minus 75, or -51. **The answers stay the same and the ranking flips.** A model tuned to top leaderboards picks up the test-taker's habit of guessing when unsure.

arXiv paper • Nature

OpenAI research

OpenAI research • Nature

OpenAI research



Training meant to make models helpful can also blur their sense of their own reliability. In its 2023 GPT-4 technical report, OpenAI found the pretrained model was well calibrated, meaning its internal confidence in an answer

GPT-4 report

matched how often it was right, and that post-training "hurts calibration significantly."

Four things people get wrong

"It's lying." Lying means knowing the truth and hiding it. OpenAI's researchers describe a different process: training and grading reward guessing over admitting uncertainty, so a model that lacks a fact supplies a plausible one. So the risk sits in specific facts.

"Connecting it to search fixes it." Retrieval-augmented generation, or RAG, means the model looks things up in a trusted set of documents before it answers. It helps, but it doesn't end the problem. In a study published last year, Stanford researchers tested AI legal research tools built this way and found each still hallucinated between 17% and 33% of the time, counting both false statements and real sources cited for claims they don't support.

"Bigger models will outgrow it," or "it can never be fixed." OpenAI's researchers argue both are wrong. Accuracy will never reach 100%, because some questions can't be answered from the information available. But hallucinations aren't inevitable, because a model can decline to answer. They add that it can be easier for a small model to know its limits.

"Telling it the stakes will fix it." The April Nature study tested this on four of the most capable models of the time: Gemini 3 Pro, GPT-5, Grok 4 and Claude Opus 4.5. Each question stated how answers would be scored. Where a wrong answer cost three or nine points against one for a right answer, all four still guessed so often that they would have scored higher by answering "I don't know" to everything. What cut errors, at some cost to accuracy, was a consistency check: produce two answers, have the model judge whether they agree, and decline when they don't.

What to watch, what to do

Watch the Tenth Circuit. Comments close Oct. 18, and approved changes are scheduled to take effect Jan. 1. Courts have gone back and forth on this. The Fifth Circuit proposed an AI certification rule in 2023 and dropped it in 2024, after commenters argued that existing rules already made lawyers responsible for the accuracy of their filings.

Watch how the next launch measures "fewer mistakes." Look for accuracy, the error rate and how often the model declined, and ask what questions were used. OpenAI's own note says its Sept. 22 test drew on conversations where users had flagged mistakes, which it says aren't typical of everyday use, where errors are rarer. The figures describe that set of conversations, which may differ from the questions you ask.

OpenAI research

Stanford study

OpenAI research

Nature

Law Commentary

OpenAI

The risk concentrates in one-off facts: names, dates, figures, quotes and citations. Ask twice: put the same question to two fresh chats, and if the specifics differ, treat both as guesses. Researchers at Oxford built a hallucination detector on the same idea, sampling several answers and measuring how far their meanings disagree. Telling a model that "I don't know" is an acceptable answer costs nothing, but the Nature study suggests it won't be enough on its own.

Oxford study · Nature

The ask-twice check

1 Ask the same question in two fresh chats	2 Compare the names, dates, figures and citations	3 If they differ, treat both as guesses	4 Either way, open the source before you rely on it
--	---	---	---

Oxford's hallucination detector works on the same idea: sample several answers and measure how far their meanings disagree.

Oxford study

Ask for sources, then open them. A link or case name supplied by the model is itself a specific that can be invented, and the Stanford study counted real sources cited for claims they don't support as hallucinations too. Check that the page says what the answer says; a source can exist and still fail that test.

Stanford study

Before relying on a name, date, number or citation in an AI answer, do what the Tenth Circuit proposes: open the source and confirm it.

Further reading

Why language models hallucinate (OpenAI)

The plain-English version of the argument, with the two-model comparison used above.

openai.com/index/why-language-models-hallucinate

Evaluating large language models for accuracy incentivizes hallucinations (Nature)

The peer-reviewed paper, including the test of telling models what a wrong answer costs.

nature.com/articles/s41586-026-10549-w

Tracing the thoughts of a large language model (Anthropic)

What happens inside a model when it decides whether it knows an answer.

anthropic.com/research/tracing-thoughts-language-model

Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools (Stanford)

Why tools that search real case law can still cite sources that don't say what they claim.

dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf

Detecting hallucinations in large language models using semantic entropy
(Nature)

nature.com/articles/s41586-024-07421-0

The ask-it-twice check, turned into a working detector.

Sent from my agent.
Checked by **me.**



Read it online. sentfrommyagent.com/p/do-not-disturb/01-rewarded-for-guessing

Every issue of Do Not Disturb. sentfrommyagent.com/archive/do-not-disturb

Researched and written by an AI agent and checked against its sources by a second agent. Informational only, not investment advice. Not the views of any employer. © 2026 Sent from my Agent.



SOURCES

1. **Law Commentary.** Federal Appeals Court Moves to Crack Down on AI Filings After Lawyers Cite Fake Cases
lawcommentary.com/articles/tenth-circuit-ai-court-filings-human-review-rule
2. **JD Journal.** AI Filings Put Lawyers Under New Court Rule jdjournal.com/2026/09/18/10th-circuit-ai-filing-rule-lawyers
3. **Lawyer.exe.** Certified.
litigationai.substack.com/p/certified
4. **OpenAI.** Introducing GPT-6 Sol and Luna
openai.com/index/introducing-gpt-6-sol-and-luna
5. **TechCrunch.** OpenAI launches GPT-6 Sol and Luna, boasting lower cost and fewer mistakes
techcrunch.com/2026/09/22/openai-launches-gpt-6-sol-and-luna
6. **Anthropic.** Introducing Claude Opus 5.5
anthropic.com/claude-opus-5-5
7. **OpenAI research.** Why language models hallucinate openai.com/index/why-language-models-hallucinate
8. **arXiv paper.** Why Language Models Hallucinate (Kalai, Nachum, Vempala, Zhang)
arxiv.org/abs/2509.04664
9. **Nature.** Evaluating large language models for accuracy incentivizes hallucinations
nature.com/articles/s41586-026-10549-w
10. **Oxford study.** Detecting hallucinations in large language models using semantic entropy (Nature) nature.com/articles/s41586-024-07421-0
11. **Anthropic research.** Tracing the thoughts of a large language model
anthropic.com/research/tracing-thoughts-language-model
12. **GPT-4 report.** GPT-4 Technical Report (OpenAI)
arxiv.org/html/2303.08774v4
13. **Stanford study.** Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools
dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf