New Scientist

Technology

AI hallucinations are getting worse – and they're here to stay

An AI leaderboard suggests the newest reasoning models used in chatbots are producing less accurate results because of higher hallucination rates. Experts say the problem is bigger than that

By Jeremy Hsu

9 May 2025

Errors tend to crop up in AI-generated content

Paul Taylor/Getty Images

AI chatbots from tech companies such as OpenAI and Google have been getting so-called reasoning upgrades over the past months – ideally to make them better at giving us answers we can trust, but recent testing suggests they are sometimes doing worse than previous models. The errors made by chatbots, known as "hallucinations", have been a problem from the start, and it is becoming clear we may never get rid of them.

Hallucination is a blanket term for certain kinds of mistakes made by the large language models (LLMs) that power systems like OpenAI's ChatGPT or Google's Gemini. It is best known as a description of the way they sometimes present false information as true. But it can also refer to an AI-generated answer that is factually accurate, but not actually relevant to the question it was asked, or fails to follow instructions in some other way.

An OpenAI technical report 🔗 https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf evaluating its latest LLMs showed that its o3 and o4-mini models, which were released in April, had significantly higher hallucination rates than the company's previous o1 model that came out in late 2024. For example, when summarising publicly available facts about people, o3 hallucinated 33 per cent of the time while o4-mini did so 48 per cent of the time. In comparison, o1 had a hallucination rate of 16 per cent.

The problem isn't limited to OpenAI. One popular leaderboard 🔗 https://github.com/vectara/hallucination-leaderboard?tab=readme-ov-file from the company Vectara that assesses hallucination rates indicates some "reasoning" models – including the DeepSeek-R1 model 🔗 /article/mg26535293-300-deepseek-has-burst-the-ai-hype-bubble-now-all-bets-are-off/ from developer DeepSeek – saw double-digit rises in hallucination rates 🔗 https://www.vectara.com/blog/deepseek-r1-hallucinates-more-than-deepseek-v3 compared with previous models from their developers. This type of model goes through multiple steps to demonstrate a line of reasoning before responding.

OpenAI says the reasoning process isn't to blame. "Hallucinations are not inherently more prevalent in reasoning models, though we are actively working to reduce the higher rates of hallucination we saw in o3 and o4-mini," says an OpenAI spokesperson. "We'll continue our research on hallucinations across all models to improve accuracy and reliability.”

Some potential applications for LLMs could be derailed by hallucination. A model that consistently states falsehoods and requires fact-checking won’t be a helpful research assistant; a paralegal-bot that cites imaginary cases will get lawyers into trouble; a customer service agent that claims outdated policies are still active will create headaches for the company.

However, AI companies initially claimed that this problem would clear up over time. Indeed, after they were first launched, models tended to hallucinate less with each update. But the high hallucination rates of recent versions are complicating that narrative – whether or not reasoning is at fault.

Vectara’s leaderboard ranks models based on their factual consistency in summarising documents they are given. This showed that “hallucination rates are almost the same for reasoning versus non-reasoning models”, at least for systems from OpenAI and Google, says Forrest Sheng Bao 🔗 https://www.linkedin.com/in/forrestbao/ at Vectara. Google didn’t provide additional comment. For the leaderboard’s purposes, the specific hallucination rate numbers are less important than the overall ranking of each model, says Bao.

But this ranking may not be the best way to compare AI models.

For one thing, it conflates different types of hallucinations. The Vectara team pointed out 🔗 https://www.vectara.com/blog/why-does-deepseek-r1-hallucinate-so-much that, although the DeepSeek-R1 model hallucinated 14.3 per cent of the time, most of these were “benign”: answers that are factually supported by logical reasoning or world knowledge, but not actually present in the original text the bot was asked to summarise. DeepSeek didn’t provide additional comment.

Another problem with this kind of ranking is that testing 🔗 /article/2478521-meta-amazon-and-google-accused-of-distorting-key-ai-rankings/ based on text summarisation “says nothing about the rate of incorrect outputs when [LLMs] are used for other tasks”, says Emily Bender 🔗 https://faculty.washington.edu/ebender/ at the University of Washington. She says the leaderboard results may not be the best way to judge this technology because LLMs aren’t designed specifically to summarise texts.

These models 🔗 /article/mg25934590-600-we-still-dont-really-understand-what-large-language-models-are/ work by repeatedly answering the question of “what is a likely next word” to formulate answers to prompts, and so they aren’t processing information in the usual sense of trying to understand what information is available in a body of text, says Bender. But many tech companies still frequently use the term “hallucinations” when describing output errors.

“‘Hallucination’ as a term is doubly problematic,” says Bender. “On the one hand, it suggests that incorrect outputs are an aberration, perhaps one that can be mitigated, whereas the rest of the time the systems are grounded, reliable and trustworthy. On the other hand, it functions to anthropomorphise the machines – hallucination refers to perceiving something that is not there [and] large language models do not perceive anything.”

Arvind Narayanan 🔗 https://www.cs.princeton.edu/~arvindn/ at Princeton University says that the issue goes beyond hallucination. Models also sometimes make other mistakes, such as drawing upon unreliable sources or using outdated information. And simply throwing more training data and computing power 🔗 /article/2449427-ais-get-worse-at-answering-simple-questions-as-they-get-bigger/ at AI hasn’t necessarily helped.

The upshot is, we may have to live with error-prone AI. Narayanan said in a social media post 🔗 https://x.com/random_walker/status/1919359709062033850 that it may be best in some cases to only use such models for tasks when fact-checking the AI answer would still be faster than doing the research yourself. But the best move may be to completely avoid relying on AI chatbots to provide factual information, says Bender.