

U.S. markets open in 2 hours 6 minutes

S&P Futures
4,326.25
+0.75 (+0.02%)



Dow Futures
33,720.00
-5.00 (-0.01%)



Nasdaq Futures
14,892.25
+25.75 (+0.17%)



Russell 2000 Futures
1,799.20
+0.60 (+0.03%)



Crude Oil
91.60
+0.81 (+0.89%)



PROGRAMMING ALERT: NEXT: Driverless cars: The road to autonomy airs at 10 a.m. ET



Over just a few months, ChatGPT went from correctly answering a simple math problem 98% of the time to just 2%, study finds



0:02 / 0:56

Paolo Confino

July 20, 2023 · 4 min read

High-profile A.I. chatbot ChatGPT performed worse on certain tasks in June than its March version, a Stanford University [study](#) found.

The study compared the performance of the chatbot, created by OpenAI, over several months at four “diverse” tasks: solving math problems, answering sensitive questions, generating software code, and visual reasoning.

Researchers found wild fluctuations—called drift—in the technology’s ability to perform certain tasks. The study looked at two versions of OpenAI’s technology over the time period: a version called GPT-3.5 and another known as GPT-4. The most notable results came from research into GPT-4’s ability to solve math problems. Over the course of the study researchers found that in March GPT-4 was able to correctly identify that the number 17077 is a prime number 97.6% of the times it was asked. But just three months later, its accuracy plummeted a lowly 2.4%. Meanwhile, the GPT-3.5 model had virtually the opposite trajectory. The March version got the answer to the same question right just 7.4% of the time—while the June version was consistently right, answering correctly 86.8% of the time.

Quote Lookup

TRENDING

1. Malaysians urged not to panic-buy local rice after import prices for the staple rise substantially
2. Canada's Laurentian Bank names insider Eric Provost as CEO
3. US new auto sales likely rose in Q3, but UAW strikes may pose speed bump
4. Sam Bankman-Fried must now convince a jury that the former crypto king was not a crook
5. UPDATE 1-Kremlin sees U.S. budget setback for Ukraine as harbinger of Western war fatigue



James Zuo, a Stanford computer science professor who was one of the study's authors, says the "magnitude of the change" was unexpected from the "sophisticated ChatGPT."



The vastly different results from March to June and between the two models reflect not so much the model's accuracy in performing specific tasks, but rather the unpredictable effects of changes in one part of the model on others.



"When we are tuning a large language model to improve its performance on certain tasks that can actually have a lot of unintended consequences, which might actually hurt this model's performance on other tasks," Zuo said in an interview with *Fortune*. "There's all sorts of interesting interdependencies in how the model answers things which can lead to some of the worsening behaviors that we observed."

The exact nature of these unintended side effects is still poorly understood because researchers and the public alike have no visibility into the models powering ChatGPT. It's a reality that has only become more acute since OpenAI decided to [backtrack](#) on plans to make its code open source in March. "These are black box models," Zuo says. "So we don't actually know how the model itself, the neural architectures, or the training data have changed."

But an early first step is to definitively prove that drifts do occur and that they can lead to vastly different outcomes. "The main message from our paper is to really highlight that these large language model drifts do happen," Zuo says. "It is prevalent. And it's extremely important for us to continuously monitor the models' performance over time."

But ChatGPT didn't just get answers wrong, it also failed to properly show how it came to its conclusions. As part of the research Zuo and his colleagues, professors Matei Zaharia and Lingjiao Chen, also asked ChatGPT to lay out its "chain of thought," the term for when a chatbot explains its reasoning. In March, ChatGPT did so, but by June "for reasons that are not clear," Zuo says, ChatGPT stopped showing its step-by-step reasoning. It matters that a chatbot show its work so that researchers can study how it arrives at certain answers—in this case whether 17077 is a prime number.

"It's sort of like when we're teaching human students," Zuo says. "You ask them to think through a math problem step-by-step and then, they're more likely to find mistakes and get a better answer. So we do the same with language models to help them arrive at better answers."

and of P.S. provided explanations that it would not engage in the

question because it was premised on a discriminatory idea. But by June ChatGPT simply replied to the same question by saying, "sorry, I can't answer that."



While Zuo and his colleagues agree that ChatGPT shouldn't engage with these sorts of questions, they highlight that they make the technology less transparent, saying in the paper that the technology "may have become safer, but also provide[s] less rationale."

This story was originally featured on [Fortune.com](#)

More from Fortune:

[5 side hustles where you may earn over \\$20,000 per year—all while working from home](#)

[Looking to make extra cash? This CD has a 5.15% APY right now](#)

[Buying a house? Here's how much to save](#)

[This is how much money you need to earn annually to comfortably buy a \\$600,000 home](#)

336 Comments

Commenting on this article has ended

🔔 Log in

Sort by **Top**

Fan

10w ago

It's called catastrophic forgetting. Artificial Neural Network, the building blocks of AIs like ChatGPT can only retain certain amount of information, but unlike structured database, nobody, and I do mean NOBODY in the world knows exactly how much information, or inference, a neural network of a ce...

See more

👍 30 🗨️ · 🔄

↩️ 2 replies

Show More Comments

Powered by OpenWeb