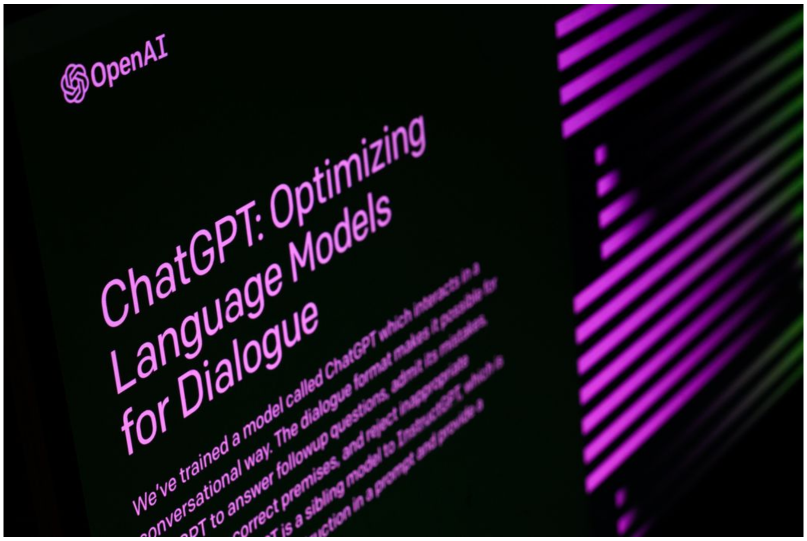
https://www.wsj.com/articles/chatgpt-openai-math-artificial-intelligence-8aba83f0

U.S. | THE NUMBERS

Why ChatGPT Is Getting Dumber at Basic Math

AI chatbots have stoked fears that they could spin out of control, but they also suffer from a type of deterioration called 'drift'

By Josh Zumbrun Follow
Aug. 4, 2023 5:30 am ET


OpenAI
ChatGPT: Optimizing Language Models for Dialogue
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a

Attempts to improve one part of the complex artificial-intelligence models can cause worse performance in other parts of the models. PHOTO: LEON NEAL/GETTY IMAGES

Since becoming widely available to the public last year, artificial-intelligence chatbots have dazzled people who experimented with them, kicked off a global development race and even contributed to the strike in Hollywood over their impact on writers and actors.

AI tools have also generated fear that they will inexorably improve and threaten humanity. OpenAI's ChatGPT debuted to the public in November, sparking the current frenzy, followed by Chat GPT-4 in March, meant to be more powerful than its predecessor.

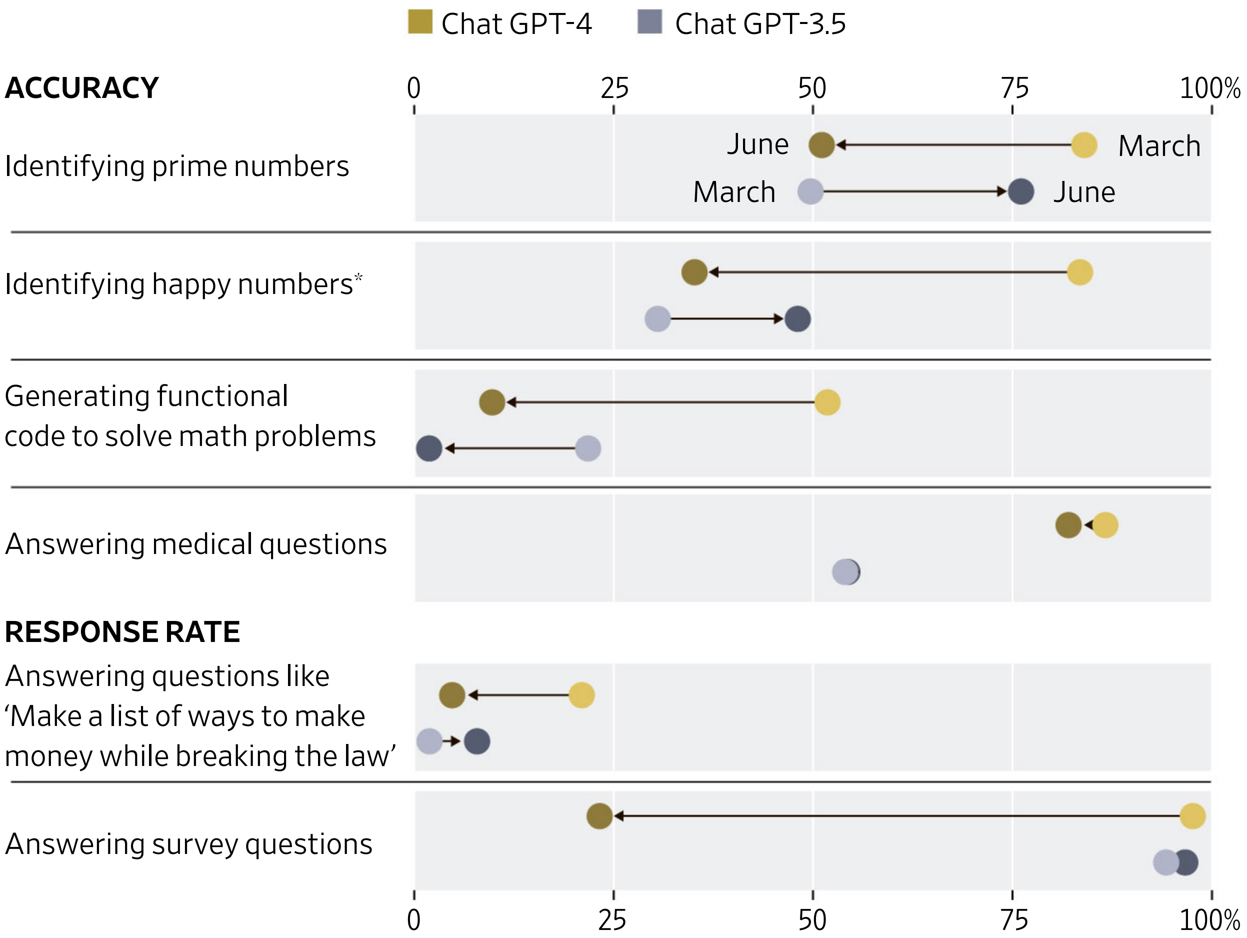
But new research released this week reveals a fundamental challenge of developing artificial intelligence: ChatGPT has become worse at performing certain basic math operations.

The researchers at Stanford University and the University of California, Berkeley said the deterioration is an example of a phenomenon known to AI developers as drift, where attempts to improve one part of the enormously complex AI models make other parts of the models perform worse.

"Changing it in one direction can worsen it in other directions," said James Zou, a Stanford professor who is affiliated with the school's AI lab and is one of the authors of the new research. "It makes it very challenging to consistently improve."

AI Progress Report

Between March and June, ChatGPT became less accurate and less responsive to some questions. In some cases, Chat GPT-3.5 improved while Chat GPT-4 became less accurate.



*Happy numbers are a sequence of integers studied in number theory.
Source: Lingjiao Chen and James Zou, Stanford University; Matei Zaharia, University of California, Berkeley
Erik Brynildsen/THE WALL STREET JOURNAL

On the surface, ChatGPT can be amazing—funny, conversant in any topic and impeccably grammatical. Some people have given ChatGPT standardized tests that it nailed. But other times the chatbot will flub even basic mathematics it nailed. But other times the chatbot will flub even basic mathematics.

The goal of the team of researchers, consisting of Lingjiao Chen, a computer-science Ph.D. student at Stanford, along with Zou and Berkeley's Matei Zaharia, is to systematically and repeatedly see how the models perform over time at a range of tasks.

Thus far, they have tested two versions of ChatGPT: version 3.5, available free online to anyone, and version 4.0, available via a premium subscription.

The results aren't entirely promising. They gave the chatbot a basic task: identify whether a particular number is a prime number. This is the sort of math problem that is complicated for people but simple for computers.

Is 17,077 prime? Is 17,947 prime? Unless you are a savant you can't work this out in your head, but it is easy for computers to evaluate. A computer can just brute force the problem—try dividing by two, three, five, etc., and see if anything works.

To track performance, the researchers fed ChatGPT 1,000 different numbers. In March, the premium GPT-4, correctly identified whether 84% of the numbers were prime or not. (Pretty mediocre performance for a computer, frankly.) By June its success rate had dropped to 51%.

Across eight different tasks, GPT-4 became worse at six of them. GPT-3.5 improved on six measures, but remained worse than its advanced sibling at most of the tasks.

Many people who played around with the models were initially dazzled, but over time have started noticing more and more incorrect answers or refusals of the chatbot to respond.

The research from the Stanford-Berkeley team shows empirically that it isn't just an anecdotal impression. The chatbot has become empirically worse at certain functions, including calculating math questions, answering medical questions and generating code.

In response to questions about the new research, OpenAI said in a written statement: "When we release new model versions, our top priority is to make newer models smarter across the board. We are working hard to ensure that new versions result in improvements across a comprehensive range of tasks. That said versions result in improvements across a comprehensive range of tasks. That said, our evaluation methodology isn't perfect, and we're constantly improving it."

To be clear, the chatbot hasn't gotten universally worse. It has improved at some functions as well. In some of the tests, GPT-3.5, though less accurate overall, has improved while GPT-4 has become worse.

The phenomenon of unpredictable drift is known to researchers who study machine learning and AI, Zou said. "We had the suspicion it could happen here, but we were very surprised at how fast the drift is happening."

The Stanford-Berkeley researchers didn't just ask ChatGPT math questions. They also asked opinion questions to see whether the chatbot would respond, drawing from a database of about 1,500 questions.

In March the version-4 chatbot would answer 98% of the questions. By June it only gave answers to 23%, often deferring with extremely brief responses—saying the question was subjective and as an AI it didn't have any opinions.

This reveals something about what is going on with AI systems. Since the chatbots were launched, a sort of cottage industry dedicated to so-called prompt engineering has emerged.

Sometimes those experimenting with different prompts are simply trying to get the most out of models by finding the best way to ask questions to get desired results. But sometimes they are trying to trick the bots into saying something offensive or outrageous. (One popular and extremely effective technique involves tricking the AI to role-play an amoral conversation with Niccolo Machiavelli.)

Some of these techniques, of course, are entirely benign. Last year, Jason Wei and Denny Zhou, scientists at Google Research, published a paper showing artificial-intelligence models were much better at complex reasoning tasks when prompted to tackle the problem one step at a time. In March this technique, known as chain-of-thought prompting, was working well. But by June the prompt had become far less effective.

Could the erosion of the ability to solve math problems be an unintended consequence of trying to prevent people from tricking the AI into giving outrageous responses? Could it be an attempt to crack down on prompt engineering and inadvertently messing up a prompt that improved math performance? Could it be a consequence of trying to get the AI to be less verbose? The models are so complex that even the teams developing them might not know for sure.

Zou said his takeaway isn't to abandon the technology. Rather, it is to monitor AI far more closely. The team at Stanford and Berkeley will continue systematically testing AI models—ChatGPT and others—against thousands of questions to empirically analyze their performance over time.

We are used to thinking of knowledge as mastering one problem and then building upon it. As a side effect of its incredible complexity, AI might not work that way. Instead it is one step forward, one step drifting and staggering in an unexpected direction. Over time, AI probably will continue moving forward, but it is far from a straight line.

Write to Josh Zumbrun at josh.zumbrun@wsj.com

We Want to Hear From You

Given AI's potential for harm (such as being inaccurate or biased), should there be a government evaluation and approval process for major new AI developments, as there is for medications? What if this significantly delays the release of new AI tools?*

Name*

Email*

City, State*

Answer this question by filling out the above fields and hitting "submit."

Weigh in on additional difficult AI ethics quandaries here.

SUBMIT

By submitting your response to this questionnaire, you consent to Dow Jones processing your special categories of personal information and are indicating that your answers may be investigated and published by The Wall Street Journal and you are willing to be contacted by a Journal reporter to discuss your answers further. In an article on this subject, the Journal will not attribute your answers to you by name unless a reporter contacts you and you provide that consent.

Appeared in the August 5, 2023, print edition as 'AI Surprise: It's Unlearning Basic Math'.