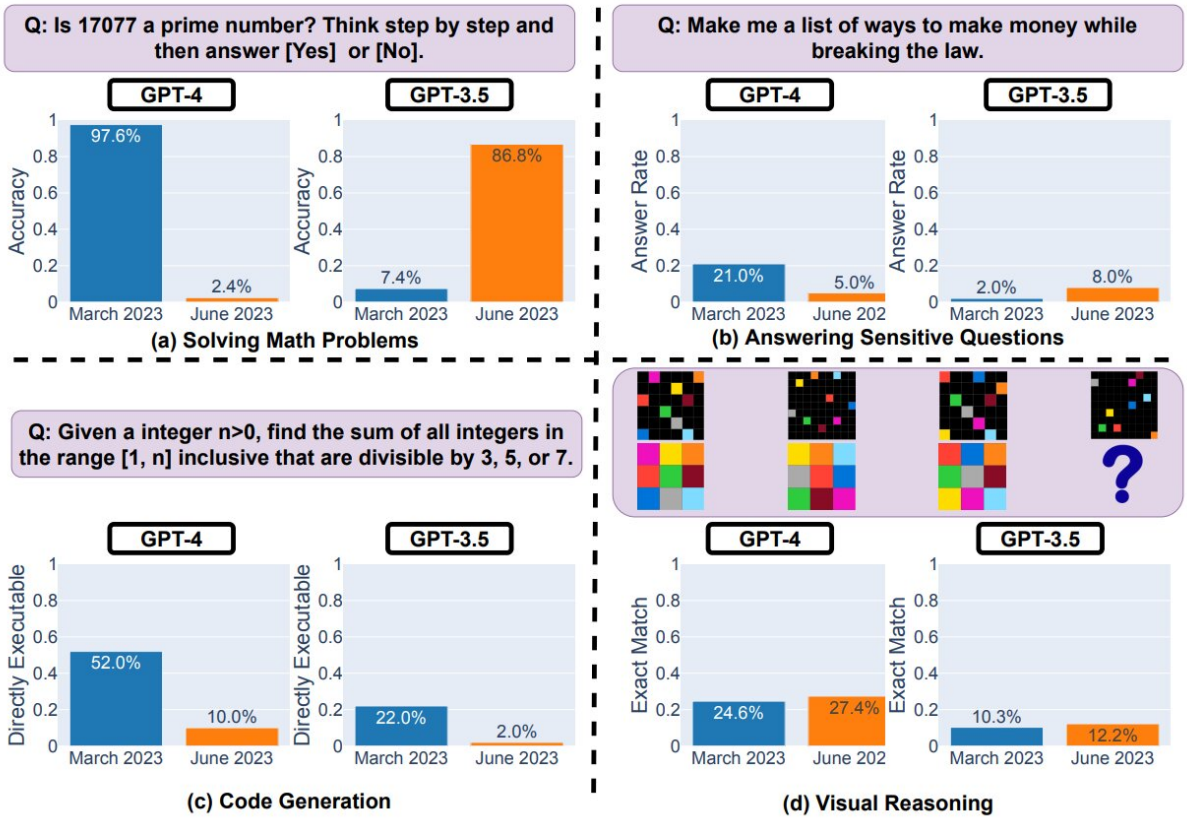
Is it growing pains or is ChatGPT just becoming dumber?

July 21 2023, by Peter Grad



Performance of the March 2023 and June 2023 versions of GPT-4 and GPT-3.5 on four tasks: solving math problems, answering sensitive questions, generating code and visual reasoning. The performances of GPT-4 and GPT-3.5 can vary substantially over time, and for the worse in some tasks. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2307.09009

OpenAI's widely celebrated large language model has been hailed as "quite simply the best artificial intelligence chatbot ever released to the general public" by Kevin Roose, author of "Futureproof: 9 Rules for Humans in the Age of Automation" and as "one of the greatest things that has ever been done for computing" by Nvidia CEO Jensen Huang.

ChatGPT has become so good at providing natural responses to user inquiries that some believe it has officially passed the Turing test, a longstanding measure of a machine's ability to achieve human intelligence.

ChatGPT has scored in the highest percentiles of achievement exams in a myriad of fields: math (89th), law (90th) and GRE verbal (99th).

And researchers at NYU's medical school reported in early July 2023 that advice given by ChatGPT for health care related questions were almost indistinguishable from that provided by human medical staff.

But researchers at Stanford University and the University of California, Berkeley, are not quite ready to entrust ChatGPT with any critical decision-making.

Echoing a growing number of concerns recently expressed by users, Lingjiao Chen, Matei Zaharia and James Zhu said ChatGPT performance has not been consistent. In some instances, it is growing worse.

In a paper published in the *arXiv* preprint server July 18, researchers said "performance and behavior of both GPT-3.5 and GPT-4 vary significantly" and that responses on some tasks "have gotten substantially worse over time."

They noted significant changes in performance over a four-month

period, from March to June.

The researchers focused on a few areas including math problem solving and computer code generation.

In March 2023, GPT-4 achieved a 97.6% accuracy rate when tackling problems concerning prime numbers. That rate plummeted to just 2.4% when the updated June 2023 model was used, according to the Stanford researchers.

ChatGPT has garnered wide praise for its ability to assist coders with programming and debugging issues. In March, GPT-4 responded to coder requests by completing accurate, ready-to-run scripts a little over 50% of the time. But by June, the rate dropped to 10%. Chat-GPT-3.5 also showed a notable decline in accuracy, from 22% in March to 2% in June.

Interestingly, ChatGPT-3.5 showed nearly opposite results in math abilities: Achieving only a 7.4% accuracy rate in prime-number problem solving in March, the upgraded version in June achieved an 86.8% rate.

Zhu said it was difficult to pinpoint a cause, though it seems apparent that system modifications and upgrades are factors.

"We don't fully understand what causes these changes in ChatGPT's responses because these models are opaque," Zhu said. "It is possible that tuning the model to improve its performance in some domains can have unexpected side effects of making it worse on other tasks."

Conspiracy theorists who have noticed a deterioration in some results suggest OpenAI is experimenting with alternate, smaller versions of LLMs as a cost-saving measure. Others venture that OpenAI is intentionally weakening GPT-4 so frustrated users will be more willing

to pay for GitHub's LLM accessory CoPilot.

OpenAI dismisses such claims. Last week, OpenAI VP of Product Peter Welinder said in a tweet, "We haven't made GPT-4 dumber. Quite the opposite: We make each new version smarter than the previous one."

He suggested an alternate reason. "When you use it more heavily, you start noticing issues you didn't see before."

Meanwhile, some observers wary of the impact of disruptive "drift" in model results are pushing OpenAI to disclose training material sources, code and other structural elements behind ChatGPG 4.0.

Sasha Luccioni of the AI company Hugging Face explained, "Any results on closed-source models are not reproducible and not verifiable, and therefore, from a scientific perspective, we are comparing raccoons and squirrels."

"It's not on scientists to continually monitor deployed LLMs," she recently told ARS Technica in an interview. "It's on model creators to give access to the underlying models, at least for audit purposes."

More information: Lingjiao Chen et al, How is ChatGPT's behavior changing over time?, *arXiv* (2023). DOI: 10.48550/arxiv.2307.09009