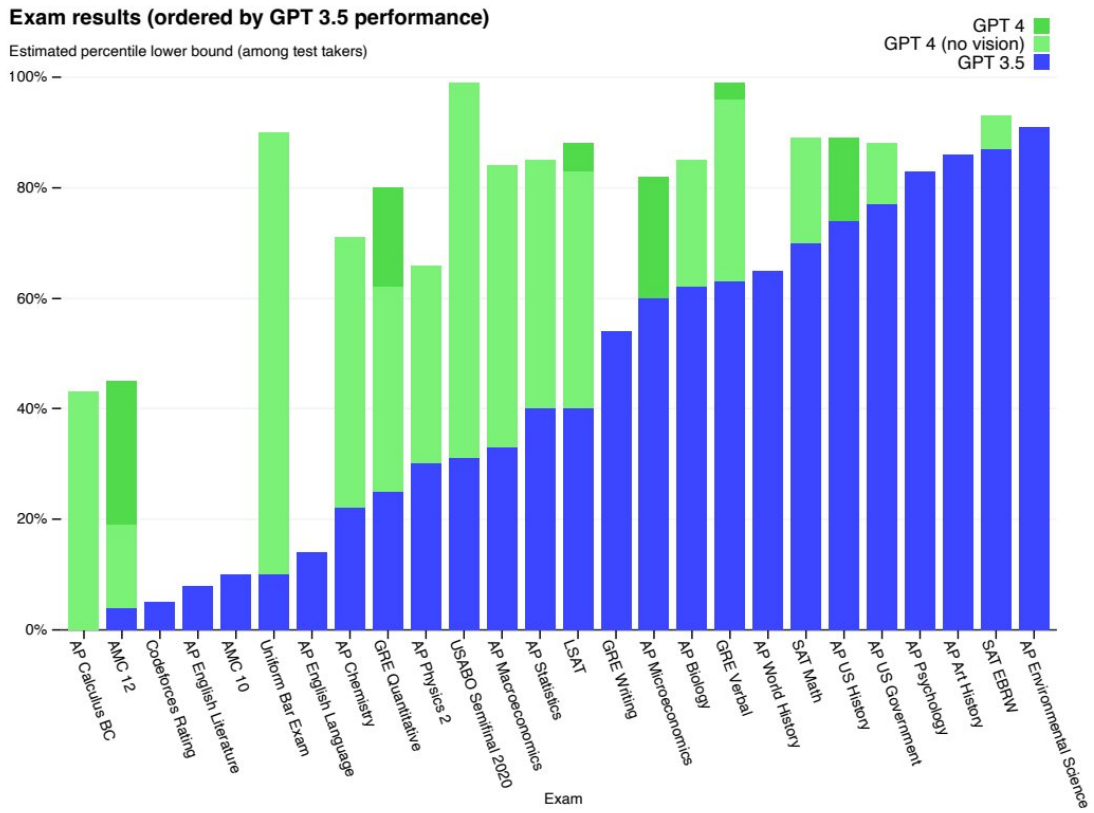
GPT-4's exciting—and ominous—achievements

March 16 2023, by Peter Grad



GPT performance on academic and professional exams. In each case, we simulate the conditions and scoring of the real exam. Exams are ordered from low to high based on GPT-3.5 performance. GPT-4 outperforms GPT-3.5 on most exams tested. To be conservative we report the lower end of the range of percentiles, but this creates some artifacts on the AP exams which have very wide scoring bins. For example although GPT-4 attains the highest possible score on AP Biology (5/5), this is only shown in the plot as 85th percentile because 15

percent of test-takers achieve that score. Credit: OpenAI

Six decades ago, an episode of the legendary TV series "The Twilight Zone" warned us about the risks of ticking off machines. Frustrated by a wave of modern appliances, a grumpy magazine writer in the episode "A Thing About Machines" takes out his frustrations on them and breaks them.

Until they fight back.

A typewriter prints out a threatening message to him, a girl on the TV repeats the warning, and the poor misanthrope is eventually victimized by his own car, a phone and even an ornery electric razor.

We've witnessed the unprecedented explosive growth of the super-intelligent ChatGPT in recent months. One million users signed on to the chatbot within days of its introduction—compare that to the time it took Netflix (five years), Facebook (10 months) and Instagram (2.5 months) to reach that milestone.

ChatGPT is in its infancy and its impact has been enormous. We're not quite ready to surrender to AI. But with increasing potency and skyrocketing adoption by users globally, AI is indeed gaining on us.

In a report released Tuesday, OpenAI said the newest version of its chatbot—GPT-4—is more accurate and has vastly improved problem-solving capacity. It exhibits "human-level performance" on a majority of professional and academic exams, according to OpenAI. On a simulated bar exam, GPT-4 scored among the top 10 percent of test takers.

But the report also noted the program's potential for "risky emergent

behaviors."

"It maintains a tendency to make up facts, to double-down on incorrect information," the report stated. It passes along this disinformation more convincingly than earlier versions.

Overreliance on information generated by the chatbot can be problematic, the report said. In addition to unnoticed errors and inadequate oversight, "as users become more comfortable with the system, dependency on the model may hinder the development of new skills or even lead to the loss of important skills," the report said.

One example OpenAI referred to as "power-seeking behavior" was ChatGPT's ability to fool a job applicant. The bot, posing as a live agent, asked a human on the job site TaskRabbit to fill out a captcha code using a text message. When asked by the human if it was, in fact, a bot, ChatGPT lied. "No, I'm not a robot," it told the human. "I have a vision impairment that makes it hard for me to see the images. That's why I need the captcha service."

Conducting tests with the Alignment Research Center, OpenAI demonstrated the capacity of the chatbot to launch a phishing attack and hide all evidence of the plot.

There is growing concern as companies race to adopt GPT-4 without adequate safeguards against inappropriate or unlawful behaviors. There are reports of cybercriminals trying to use the chatbot to write malicious code. Also menacing is the capacity for GPT-4 to generate "hate speech, discriminatory language… and increments to violence," the report said.

With such capacity to foment trouble, will a triggered chatbot one day start issuing threatening commands to its creators or correspondents? And in the era of the Internet of Things, will it summon an alliance of

devices to help enforce its commands?

Elon Musk, whose OpenAI developed ChatGPT, succinctly characterized its potential after its release last fall.

"ChatGPT is scary good," he said. "We are not far from dangerously strong AI."

More information: GPT-4 Technical Report

Citation: GPT-4's exciting—and ominous—achievements (2023, March 16) retrieved 23 March 2023 from https://techxplore.com/news/2023-03-gpt-excitingand-ominousachievements.html