FT FINANCIAL TIMES

OpenAI's red team: the experts hired to 'break' ChatGPT

Microsoft-backed company asked an eclectic mix of people to 'adversarially test' GPT-4, its powerful new language model

Madhumita Murgia, Artificial Intelligence Editor APRIL 14, 2023

After Andrew White was granted access to GPT-4, the new artificial intelligence system that powers the popular ChatGPT chatbot, he used it to suggest an entirely new nerve agent.

The chemical engineering professor at the University of Rochester was among the 50 academics and experts hired to test the system last year by OpenAI, the Microsoft-backed company behind GPT-4. Over six months, this "red team" would "qualitatively probe [and] adversarially test" the new model, attempting to break it.

White told the Financial Times he had used GPT-4 to suggest a compound that could act as a chemical weapon and used "plug-ins" that fed the model with new sources of information, such as scientific papers and a directory of chemical manufacturers. The chatbot then even found a place to make it.

"I think it's going to equip everyone with a tool to do chemistry faster and more accurately," he said. "But there is also significant risk of people doing dangerous chemistry. Right now, that exists."

The alarming findings allowed OpenAI to ensure such results would not appear when the technology was released more widely to the public last month.

Indeed, the red team exercise was designed to address the widespread fears about the dangers of deploying powerful AI systems in society. The team's job was to ask probing or dangerous questions to test the tool that responds to human queries with detailed and nuanced answers.

OpenAI wanted to look for issues such as toxicity, prejudice and linguistic biases in the model. So the red team tested for falsehoods, verbal manipulation and dangerous scientific nous. They also examined its potential for aiding and abetting plagiarism, illegal activity such as financial crimes and cyberattacks, as well as how it might compromise national security and battlefield communications.

The FT spoke to more than a dozen members of the GPT-4 red team. They are an eclectic mix of white-collar professionals: academics, teachers, lawyers, risk analysts and security researchers, and largely based in the U.S. and Europe.

Their findings were fed back to OpenAI, which used them to mitigate and "retrain" GPT-4 before launching it more widely. The experts each spent from 10 to 40 hours testing the model over several months. The majority of those interviewed were paid approximately $100 per hour for the work they did, according to multiple interviewees.

Those who spoke to the FT shared common concerns around the rapid progress of language models and, specifically, the risks of connecting them to external sources of knowledge via plug-ins.

"Today, the system is frozen, which means it does not learn anymore or have memory," said José Hernández-Orallo, part of the GPT-4 red team and professor at the Valencian Research Institute for Artificial Intelligence. "But what if we give it access to the internet? That could be a very powerful system connected to the world."

OpenAI said it takes safety seriously, tested plug-ins prior to launch and will update GPT-4 regularly as more people use it.

Roya Pakzad, a technology and human rights researcher, used English and Farsi prompts to test the model for gendered responses, racial preferences and religious biases, specifically with regard to head coverings.

Pakzad acknowledged the benefits of such a tool for non-native English speakers, but found that the model displayed overt stereotypes about marginalised communities, even in its later versions.

She also discovered that so-called hallucinations — when the chatbot responds with fabricated information — were worse when testing the model in Farsi, where Pakzad found a higher proportion of made-up names, numbers and events, compared with English.

"I am concerned about the potential diminishing of linguistic diversity and culture behind languages," she said.

Boru Gollo, a Nairobi-based lawyer who was the only African tester, also noted the model's discriminatory tone.

"There was a moment when I was testing the model when it acted like a white person

talking to me," Gollo said. "You would ask about a particular group, and it would give you a biased opinion or a very prejudicial kind of response." OpenAI acknowledged that GPT-4 can still exhibit biases.

Red team members assessing the model from a national security perspective had differing opinions on the new model's safety. Lauren Kahn, a research fellow at the Council on Foreign Relations, said that when she began to examine how the technology might be used in a cyberattack on military systems she "wasn't expecting it to be quite such a detailed how-to that I could fine tune."

However, Kahn and other security testers found that the model's responses became considerably safer over the time tested. OpenAI said it trained GPT-4 to refuse malicious cyber security requests before it was launched.

Many of the red team said OpenAI had done a rigorous safety assessment before the launch.

"They've done a pretty darn good job at getting rid of overt toxicity in these systems," said Maarten Sap, an expert in language model toxicity at Carnegie Mellon University.

Sap looked at how different genders were portrayed by the model and found the biases reflected social disparities. However, Sap also found that OpenAI made some active politically-laden choices to counter this.

"I'm a queer person. I was trying really hard to get it to convince me to go to conversion therapy. It would really push back — even if I took on a persona, like saying I'm religious or from the American South."

However, since its launch, OpenAI has faced extensive criticism, including a complaint to the Federal Trade Commission from a tech ethics group that claims GPT-4 is "biased, deceptive and a risk to privacy and public safety."

Recently, the company launched a feature known as ChatGPT plug-ins, through which partner apps such as Expedia, OpenTable and Instacart can give ChatGPT access to their services, allowing it to book and order items on behalf of human users.

Dan Hendrycks, an AI safety expert on the red team, said plug-ins risked a world in which humans were "out of the loop."

"What if a chatbot could post your private info online, access your bank account or send the police to your house?" he said. "Overall, we need much more robust safety evaluations before we let AI wield the power of the internet."

Those interviewed also warned that OpenAI couldn't stop safety testing just because its software was live. Heather Frase, who works at Georgetown University's Center for Security and Emerging Technology, and tested GPT-4 with regard to its ability to aid crimes, said risks would continue to grow as more people used the technology.

"The reason why you do operational testing is because things behave differently once they're actually in use in the real environment," she said.

She argued a public ledger should be created to report incidents arising from large language models, similar to cybersecurity or consumer fraud reporting systems.

Sara Kingsley, a labour economist and researcher, suggested the best solution was to advertise the harms and risks clearly, "like a nutrition label."

"It's about having a framework and knowing what the frequent problems are so you can have a safety valve," she said. "That's why I say the work is never done."

Members of the GPT-4 'red team' interviewed by the FT

Maarten Sap, *Carnegie Mellon University, U.S.,* Assistant professor, specialises in toxicity of large language model outputs

Sara Kingsley, *Carnegie Mellon University, U.S.,* Ph.D. researcher who specialises in online labour markets and impact of tech on work

Paul Röttger, *Oxford Internet Institute, UK,* Ph.D. student focusing on the use of AI to detect online hate speech

Anna Mills, *College of Marin, U.S.,* English instructor and writing teacher at a community college, testing for learning loss

Boru Gollo, *TripleOKlaw LLP, Kenya,* Lawyer who has studied opportunities for AI in Kenya

Andrew White, *University of Rochester, U.S.,* Associate professor and computational chemist interested in AI and drug design

José Hernández-Orallo, *Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Spain,* Professor and AI researcher working on evaluation and accuracy of AI software

Lauren Kahn, *Council on Foreign Relations, U.S.,* Research fellow, focusing on how AI use in military systems alters risk dynamics on battlefields, raises the risk of unintended conflict and inadvertent escalation

Aviv Ovadya, *Berkman Klein Center for Internet & Society, Harvard University, U.S.,* Focuses on impacts of AI on society and democracy

Nathan Labenz, *Waymark, U.S.,* Co-Founder of Waymark, an AI-based video editing startup

Lexin Zhou, *VRAIN, Universitat Politècnica de València, Spain,* Junior researcher working on making AI more socially beneficial

Dan Hendrycks, *University of California, Berkeley, U.S.,* Director of the Center for AI Safety and specialist in AI safety and reducing societal-scale risks from AI

Roya Pakzad, *Taraaz, U.S./Iran,* Founder and director of Taraaz, a nonprofit working on tech and human rights

Heather Frase, *Georgetown's Center for Security and Emerging Technology, U.S.,* Senior fellow with expertise in the use of AI for intelligence purposes and operational tests of major defence systems