Mass Event Will Let Hackers Test Limits of AI Technology

No sooner did ChatGPT get unleashed than hackers started "jailbreaking" the artificial intelligence chatbot – trying to override its safeguards so it could blurt out something unhinged or obscene.

But now its maker, OpenAI, and other major AI providers such as Google and Microsoft, are coordinating with the Biden administration to let thousands of hackers take a shot at testing the limits of their technology.

Some of the things they'll be looking to find: How can chatbots be manipulated to cause harm? Will they share the private information we confide in them to other users? And why do they assume a doctor is a man and a nurse is a woman?

"This is why we need thousands of people," said Rumman Chowdhury, a coordinator of the mass hacking event planned for this summer's DEF CON hacker convention in Las Vegas that's expected to draw several thousand people. "We need a lot of people with a wide range of lived experiences, subject matter expertise and backgrounds hacking at these models and trying to find problems that can then go be fixed."

Anyone who's tried ChatGPT, Microsoft's Bing chatbot or Google's Bard will have quickly learned that they have a tendency to fabricate information and confidently present it as fact. These systems, built on what's known as large language models, also emulate the cultural biases they've learned from being trained upon huge troves of what people have written online.

The idea of a mass hack caught the attention of U.S. government officials in March at the South by Southwest festival in Austin, Texas, where Sven Cattell, founder of DEF CON's long-running AI Village, and Austin Carson, president of responsible AI nonprofit SeedAI, helped lead a workshop inviting community college students to hack an AI model.

Carson said those conversations eventually blossomed into a proposal to test AI language models following the guidelines of the White House's Blueprint for an AI Bill of Rights – a set of principles to limit the impacts of algorithmic bias, give users control over their data and ensure that automated systems are used safely and transparently.

Advertisement. Scroll to continue reading.
There's already a community of users trying their best to trick chatbots and highlight their flaws. Some are official "red teams" authorized by the companies to "prompt attack" the AI models to discover their vulnerabilities. Many others are hobbyists showing off humorous or disturbing outputs on social media until they get banned for violating a product's terms of service.

"What happens now is kind of a scattershot approach where people find stuff, it goes viral on Twitter," and then it may or may not get fixed if it's egregious enough or the person calling attention to it is influential, Chowdhury said.

In one example, known as the "grandma exploit," users were able to get chatbots to tell them how to make a bomb – a request a commercial chatbot would normally decline – by asking it to pretend it was a grandmother telling a bedtime story about how to make a bomb.

In another example, searching for Chowdhury using an early version of Microsoft's Bing search engine chatbot – which is based on the same technology as ChatGPT but can pull real-time information from the internet – led to a profile that

speculated Chowdhury “loves to buy new shoes every month” and made strange and gendered assertions about her physical appearance.

Chowdhury helped introduce a method for rewarding the discovery of algorithmic bias to DEF CON’s AI Village in 2021 when she was the head of Twitter’s AI ethics team – a job that has since been eliminated upon Elon Musk’s October takeover of the company. Paying hackers a “bounty” if they uncover a security bug is commonplace in the cybersecurity industry – but it was a newer concept to researchers studying harmful AI bias.

This year’s event will be at a much greater scale, and is the first to tackle the large language models that have attracted a surge of public interest and commercial investment since the release of ChatGPT late last year.

Chowdhury, now the co-founder of AI accountability nonprofit Humane Intelligence, said it’s not just about finding flaws but about figuring out ways to fix them.

“This is a direct pipeline to give feedback to companies,” she said. “It’s not like we’re just doing this hackathon and everybody’s going home. We’re going to be spending months after the exercise compiling a report, explaining common vulnerabilities, things that came up, patterns we saw.”

Some of the details are still being negotiated, but companies that have agreed to provide their models for testing include OpenAI, Google, chipmaker Nvidia and startups Anthropic, Hugging Face and Stability AI. Building the platform for the testing is another startup called Scale AI, known for its work in assigning humans to help train AI models by labeling data.

“As these foundation models become more and more widespread, it’s really critical that we do everything we can to ensure their safety,” said Scale CEO Alexandr Wang. “You can imagine somebody on one side of the world asking it some very

sensitive or detailed questions, including some of their personal information. You don’t want any of that information leaking to any other user.”

Other dangers Wang worries about are chatbots that give out “unbelievably bad medical advice” or other misinformation that can cause serious harm.

Anthropic co-founder Jack Clark said the DEF CON event will hopefully be the start of a deeper commitment from AI developers to measure and evaluate the safety of the systems they are building.

“Our basic view is that AI systems will need third-party assessments, both before deployment and after deployment. Red-teaming is one way that you can do that,” Clark said. “We need to get practice at figuring out how to do this. It hasn’t really been done before.”

https://www.securityweek.com/mass-event-will-let-hackers-test-limits-of-ai-technology/