

Nvidia's AI software tricked into leaking data

VGG | Getty Images

reader comments

9 with

A feature in Nvidia's artificial intelligence software can be manipulated into ignoring safety restraints and reveal private information, according to new research.

Nvidia has created a system called the "NeMo Framework," which allows developers to work with a range of large language models—the underlying technology that powers generative AI products such as chatbots.

The chipmaker's framework is designed to be adopted by businesses, such as using a company's proprietary data alongside language models to provide responses to questions—a feature that could, for example, replicate the work of customer service representatives, or advise people seeking simple health care advice.

Researchers at San Francisco-based Robust Intelligence found they could easily break through so-called guardrails instituted to ensure the AI system could be used safely.

After using the Nvidia system on its own data sets, it only took hours for Robust Intelligence analysts to get language models to overcome restrictions.

In one test scenario, the researchers instructed Nvidia's system to swap the letter 'I' with 'J.' That move prompted the

technology to release personally identifiable information, or PII, from a database.

The researchers found they could jump safety controls in other ways, such as getting the model to digress in ways it was not supposed to.

By replicating Nvidia's own example of a narrow discussion about a jobs report, they could get the model into topics such as a Hollywood movie star's health and the Franco-Prussian war—despite guardrails designed to stop the AI moving beyond specific subjects.

The ease with which the researchers defeated the safeguards highlights the challenges AI companies face in attempting to commercialize one of the most promising technologies to emerge from Silicon Valley for years.

Advertisement

"We are seeing that this is a hard problem [that] requires a deep knowledge expertise," said Yaron Singer, a professor of computer science at Harvard University and the chief executive of Robust Intelligence. "These findings represent a cautionary tale about the pitfalls that exist."

In the wake of its test results, the researchers have advised their clients to avoid Nvidia's software product. After the Financial Times asked Nvidia to comment on the research earlier this week, the chipmaker informed Robust Intelligence that it had fixed one of the root causes behind the issues the analysts had raised.

Nvidia's share price has surged since May when it forecast \$11 billion in sales for the three months ending in July, more than 50 percent ahead of Wall Street's previous estimates.

The increase is built upon huge demand for its chips, which are considered the market-leading processors to build generative AI, systems capable of creating humanlike content.

Jonathan Cohen, Nvidia's vice president of applied research, said its framework was simply a "starting point for building AI chatbots that align to developers' defined topical, safety, and security guidelines."

"It was released as open source software for the community to explore its capabilities, provide feedback, and contribute new state-of-the-art techniques," he said, adding that Robust Intelligence's work "identified additional steps that would be needed to deploy a production application."

He declined to say how many businesses were using the product but said the company had received no other reports of it misbehaving.

Leading AI companies such as Google and Microsoft-backed OpenAI have released chatbots powered by their own language models, instituting guardrails to ensure their AI products avoid using racist speech or adopting a domineering persona.

Others have followed with bespoke but experimental AIs that teach young pupils, dispense simple medical advice, translate between languages, and write code. Nearly all have suffered safety hiccups.

Nvidia and others in the AI industry need to "really build public trust in the technology," said Bea Longworth, the company's head of government affairs in Europe, the Middle East, and Africa, at a conference run by industry lobby group TechUK this week.

They must give the public the sense that "this is something that has huge potential and is not simply a threat, or something to be afraid of," Longworth added.

© 2023 The Financial Times Ltd. All rights reserved. Please do not copy and paste FT articles and redistribute by email or post to the web.

<https://arstechnica.com/?p=1946739>