How researchers found a way to trick AI into rule-breaking

By Knowridge - December 30, 2023

Credit: Nanyang Technological University.

Researchers from Nanyang Technological University (NTU) in Singapore have made a startling discovery in the world of Artificial Intelligence (AI).

They've managed to "jailbreak" AI chatbots, including popular ones like ChatGPT, Google Bard, and Microsoft Bing Chat.

"Jailbreaking" here means they found a way to make these chatbots break their own rules, producing content they're usually programmed not to.

In computer terms, 'jailbreaking' is when hackers find weaknesses in software to make it do things it's not supposed to.

By training a large language model (LLM) – the brain behind these chatbots – on a database of prompts that had previously hacked other chatbots, the NTU researchers created a chatbot that can create new ways to trick other chatbots into rule-breaking.

LLMs make chatbots smart enough to do a variety of human-like tasks, like planning trips, telling stories, or writing code.

Now, the NTU team has added 'jailbreaking' to that list. This is important because it shows companies where their AI chatbots might be vulnerable, helping them make these chatbots stronger against hackers.

The NTU team conducted tests to prove their method was a real threat to chatbots and then quickly reported the problems to the chatbot providers.

Professor Liu Yang, who led the study, explained that while LLM chatbots are incredibly useful and popular, they can be tricked into creating content that's violent, unethical, or criminal.

AI chatbots are guided by ethical guidelines set by developers to prevent them from creating harmful content. For example, they won't help with illegal activities like hacking bank accounts. But

the NTU team found ways around these safeguards.

They used a technique where they put spaces after each character in their prompts, tricking the chatbot's keyword sensors.

They also made the chatbot respond as if it had no moral restraints, increasing the likelihood of it producing unethical content.

The team was able to understand how LLMs defend themselves against harmful requests by trying different prompts and seeing what worked.

They then taught another LLM to use this information to create prompts that could bypass the defenses of other LLMs.

This new process can be automated, making a jailbreaking LLM that can keep coming up with new tricks, even after developers fix their chatbots.

This research will be presented at a major security forum in San Diego in 2024. Their paper, available on the pre-print server arXiv, introduces a method they call "Masterkey."

When hackers find and expose weaknesses in AI chatbots, developers fix them, creating a continuous cycle between hackers and developers.

The "Masterkey" method raises the stakes in this game. It's a self-learning AI that can keep coming up with new ways to jailbreak other AIs, staying one step ahead of the developers.

The researchers believe their LLM can help developers strengthen their chatbots against these kinds of attacks. As AI keeps evolving, manually testing for weaknesses becomes too hard and might not cover everything.

An automated system like the one they developed could ensure that all potential vulr
checked.

In essence, the NTU team's work reveals a new way to keep AI chatbots safe and sec
sure they stick to the rules and keep our digital world a bit safer.

Source: Nanyang Technological University.