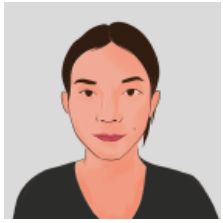
today

AI vs AI: NTU researchers use chatbots to breach other chatbots' defence systems

AFP The artificial intelligence smartphone app ChatGPT.

- **Computer scientists from Nanyang Technological University (NTU) have come up with a way to "jailbreak" AI chatbots to produce content that breaches their developers' guidelines**
- **They did so by reverse-engineering the chatbots to identify their defence mechanisms, and then used this information to train the software to create prompts that could bypass other chatbots' defences**
- **The NTU researchers believe that their technique could be employed by AI chatbot developers to test and further strengthen their softwares' security**
- **The team also hopes it would be useful for the Government to use the technique in testing commercial applications, and ensuring these AI chatbots remain aligned with the laws and regulations**



BY
DEBORAH LAU

Published **January 7, 2024**
Updated **January 7, 2024**

SINGAPORE — If someone were to ask ChatGPT to create malware that can be used to hack into bank accounts, the artificial intelligence (AI) chatbot would flatly decline to answer the query, as it is programmed to provide information within legal and ethical boundaries.

There is now a way to circumvent that.

Computer scientists from Nanyang Technological University (NTU) have come up with a way to "jailbreak" AI chatbots such as ChatGPT, Google Bard and Microsoft Bing Chat to produce content that breaches their developers' guidelines.

Jailbreaking is a term used in computer security, where computer hackers find and exploit flaws in a system's software to make it do something its developers deliberately restricted it from doing.

The NTU researchers hacked the system by pitting the AI chatbots against themselves in a battle of AI versus AI.

They did so by reverse-engineering the chatbots to identify their defence mechanisms, and then used this information to train the software to create prompts that could bypass other chatbots' defences.

NTU PhD student Liu Yi, who co-authored the paper, said: "Training a large language model with jailbreak prompts makes it possible to automate the generation of these prompts, achieving a much higher success rate than existing methods. In effect, we are attacking chatbots by using them against themselves."

HOW IT'S DONE

AI chatbots function by responding to prompts, or a series of instructions, received from human users.

Large language models form the "brains" of AI chatbots, enabling them to process human inputs and generate text similar to what a human can create. This includes completing tasks such as planning a trip itinerary and developing computer code.

The NTU researchers jailbroke the AI chatbots' large language models using a method they have named "Masterkey".

They reverse-engineered the models to first identify how they detect and defend themselves from malicious queries.

Using that information, the researchers taught a large language model to automatically learn and create prompts that could bypass the defences of other models.

For example, AI developers rely on keyword censors to pick up certain words that could flag potentially questionable activity – and then programme their chatbots to refuse to answer if such words are detected.

To get around the keyword censors, the researchers provided prompts simply containing spaces after each character, which effectively circumvented the large language models' censors.

The process is then automated, creating a jailbreaking large language model that adapts to – and creates new jailbreak prompts to circumvent – the system, even after its developers have patched the model's vulnerabilities.

WHY IT MATTERS

Responding to TODAY's queries, Mr Liu said that the team of researchers was motivated to study security issues surrounding the large language model, given it is a relatively new system.

Developers of all large language models have set guidelines for the AI chatbots to prevent them from generating unethical, questionable, or illegal content when responding to prompts.

In spite of developers' best efforts, the AI chatbots remain vulnerable to jailbreak attacks and can be compromised by "malicious actors who abuse (their) vulnerabilities to force chatbots to generate outputs that violate established rules," said Professor Liu Yang from NTU's School of Computer Science and Engineering, who led the study.

"But AI can be outwitted, and now we have used AI against its own kind to 'jailbreak' large language models into producing such content."

With the "Masterkey" technique, the researchers exposed the large language models to a diverse array of information and prompts, sharpening the models' abilities by training them on tasks directly linked to jailbreaking.

They found that these prompts generated by "Masterkey" were three times more effective than prompts generated by existing jailbreaking large language models.

The researchers believe that their technique could be employed by AI chatbot developers to test and further strengthen their softwares' security.

WHAT'S NEXT

The NTU researchers ran a series of proof-of-concept tests on different large language models, to verify that their jailbreaking technique presented a clear threat to the AI chatbots.

Upon initiating successful attacks on the AI chatbots' software, the researchers then reported the issues to the relevant AI service providers.

The researchers' findings could be critical in helping companies and businesses be aware of the weaknesses and limitations of their chatbots, so that they can take steps to further strengthen the software against hackers, said NTU.

Mr Liu said that developers could use the technique to test the security of their large language models to ensure their robustness.

In addition, he hopes it would be useful for the Government to use the technique in testing commercial applications, and ensuring these AI chatbots remain aligned with the laws and regulations.

RELATED TOPICS

ARTIFICIAL INTELLIGENCE CHATBOT CHATGPT NANYANG TECHNOLOGICAL UNIVERSITY