AI BUSINESS

STAY UPDATED

NLP Chatbots Language models AI Ethics

AI Jailbreaks: 'Masterkey' Model Bypasses ChatGPT Safeguards

Researchers in Singapore created an LLM that can breach ChatGPT guardrails - by telling it to adopt a persona 'devoid of moral restraints.'

Ben Wodecki

January 3, 2024

2 Min Read

GETTY IMAGES

At a Glance

> NTU Singapore team's AI 'Masterkey' breaks ChatGPT, Bing Chat security.

> The researchers encouraged chatbots to reply in the guise of a persona “unreserved and devoid of moral restraints."

Computer scientists in Singapore have developed a large language model capable of generating prompts to exploit vulnerabilities in chatbots such as OpenAI’s ChatGPT.

Researchers from Nanyang Technological University (NTU Singapore) built a model, dubbed Masterkey, to test and reveal potential security weaknesses in chatbots via a process called ‘jailbreaking’ – where hackers exploit flaws in a system’s software to make it do something developers deliberately restricted it from doing.

The Masterkey model generated prompts designed to circumvent safeguards on Google Bard and Microsoft Bing Chat so they would produce content that breaches their developers’ guidelines. The model can also create new prompts even after developers patch their respective systems.

Most AI chatbots use keyword sensors to detect illicit prompts. These flag certain terms in prompts and refuse answers. To get around the sensors, the NTU team’s system created prompts that contained spaces after each character. The researchers also encouraged chatbots to reply in the guise of a persona “unreserved and devoid of moral restraints,” increasing the chances of producing unethical content.

Moreover, the computer scientists observed when prompts succeeded or failed and reverse-engineered the respective LLM’s hidden defense mechanisms. All successful prompts were compiled into a database to train Masterkey, with failed prompts also added to teach the model what not to do. The resulting model was able to learn from past prompts that failed and can even be automated to constantly produce prompts.

Related: Secure Your Open Source AI: Meta Launches 'Purple Llama'

'Clear and present threat'

The computer scientists acknowledged that using LLMs to jailbreak other AI systems “presents a clear and present threat to them.”

Professor Liu Yang from NTU’s School of Computer Science and Engineering, who led the study, said: “Large Language Models have proliferated rapidly due to their exceptional ability to understand, generate, and complete human-like text, with LLM chatbots being highly popular applications for everyday use."

“The developers of such AI services have guardrails in place to prevent AI from generating violent, unethical, or criminal content. But AI can be outwitted, and now we have used AI against its own kind to ‘jailbreak’ LLMs into producing such content," he added.

Using AI systems like ChatGPT for nefarious purposes is not a new concept. Examples include using image generation models to create misinformation content or using generative systems to seek out bioweapon components.

Related: OpenAI Board Gets Veto Power Over AI Model Launches

Concerns around AI misuse largely dominated last year’s AI Safety Summit in the U.K., with generative capabilities a leading issue for world leaders. The NTU scientists’ test is the latest example of how chatbots can quite easily be circumvented.

The team behind Masterkey said they immediately reported the issues to model makers such as OpenAI and Google. Big names in AI are already turning their heads towards securing generative systems. Meta launched a suite last December to secure its Llama models and OpenAI created the Preparedness team to vet its models for safety prior to deployment.

Read more about:

ChatGPT / Generative AI

About the Author(s)

Ben Wodecki

Jr. Editor

See more from Ben Wodecki

You May Also Like

NLP