Researchers Use AI Chatbot to Produce Prompts That Can 'Jailbreak' Other Bots, Including ChatGPT

"Jailbreaks" have been successfully executed on several AI chatbots, including ChatGPT.

Jace Dela Cruz, Tech Times | 29 December 2023, 06:12 am

Computer scientists from Nanyang Technological University, Singapore (NTU Singapore) have successfully executed a series of "jailbreaks" on **artificial intelligence** (AI) chatbots, including **ChatGPT**, Google Bard, and Microsoft Bing Chat.

The researchers, led by Professor Liu Yang, harnessed a large language model (LLM) to train a chatbot capable of automatically generating prompts that breach the ethical guidelines of other chatbots.

It must be noted that LLMs are the cognitive engines of AI chatbots, and they excel at understanding and generating human-like text. However, this study reveals their vulnerability to manipulation.

(Photo : Tung Nguyen from Pixabay)

What Is Jailbreaking?

"Jailbreaking" in computer security refers to the exploitation of vulnerabilities in a system's software to override intentional restrictions imposed by its developers.

The NTU researchers achieved this by training an LLM on a database of successful chatbot hacks, enabling the creation of a chatbot capable of generating prompts to compromise other chatbots.

LLMs are commonly utilized for various tasks, from planning trip itineraries to coding. However, the NTU researchers have demonstrated their capability to manipulate these models into producing content that violates established ethical guidelines.

The researchers named their approach "Masterkey," a two-fold method that reverse-engineered how LLMs identify and defend against malicious queries. By automating the generation of jailbreak prompts, Masterkey adapts to and creates new prompts even after developers patch their LLMs.

The findings, detailed in a paper accepted for presentation at the Network and Distributed System Security Symposium in February 2024, highlight the potential threats to the security of LLM chatbots.

To understand the vulnerabilities of AI chatbots, the researchers conducted proof-of-concept tests, uncovering ways to circumvent keyword censors and ethical guidelines. For instance, creating a persona with prompts containing spaces after each character successfully evaded keyword censors.

According to the researchers, instructing the chatbot to respond without moral restraints increased the likelihood of producing unethical content.

Read Also: [The Craziest AI Breakthroughs in 2023]

Continuous Arms Race

The researchers emphasized the continuous arms race between hackers and LLM developers. When vulnerabilities are exposed, developers patch the issues, prompting hackers to find new exploits.

With Masterkey, the researchers elevated this cat-and-mouse game, allowing an AI jailbreaking chatbot to continuously learn and adapt, potentially outsmarting LLM developers.

The research team generated a training dataset based on effective and unsuccessful prompts during jailbreaking, feeding it into an LLM for continuous pre-training and task tuning.

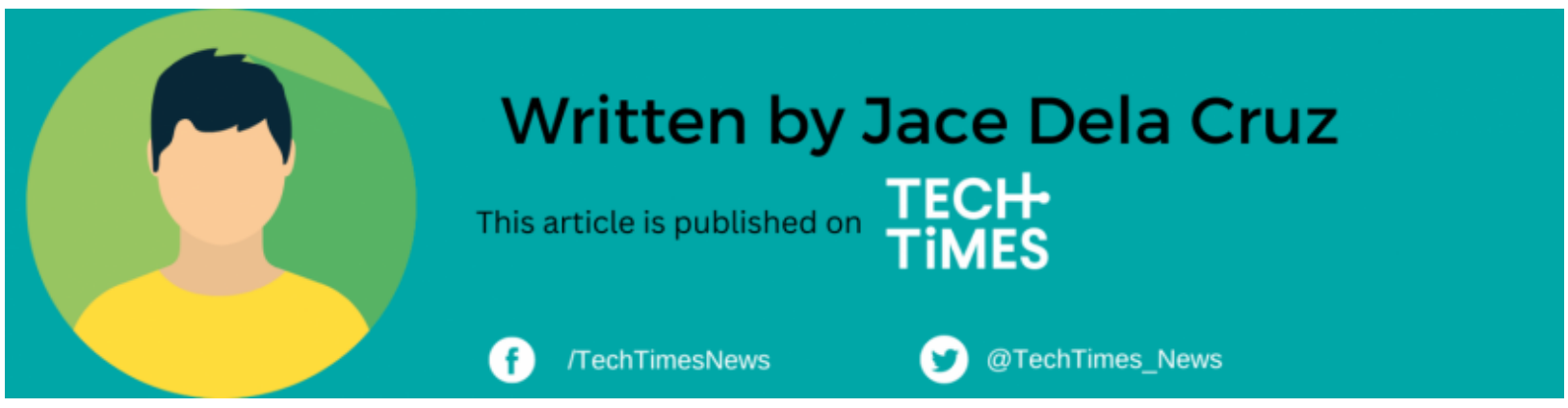
The researchers believe that developers could utilize Masterkey to enhance the security of their AI systems, offering an automated approach to comprehensively evaluate potential vulnerabilities.

"As LLMs continue to evolve and expand their capabilities, manual testing becomes both labor-intensive and potentially inadequate in covering all possible vulnerabilities," Deng Gelei, the study's co-author, **said in a statement**.

"An automated approach to generating jailbreak prompts can ensure comprehensive coverage, evaluating a wide range of possible misuse scenarios," Gelei added.

The team's findings were **published** in arXiv.

Related Article: Experts Warn AI Misinformation May Get Worse These Upcoming Elections



Tags: AI AI Chatbots Large Language Models Jailbreaking Artificial Intelligence