THE DEBRIEF

MJ BANIAS ·JANUARY 3, 2024

New research has revealed the results of pitting a specialized AI system against multiple common Large Language Model (LLM) chatbots like ChatGPT and Bard, in an attempt to break down their defense mechanisms.

In their recent study, a collective of researchers from the Nanyang Technological University (NTU) in Singapore, the University of New South Wales, the Huazhong University of Science and Technology, and Virginia Tech tested and subsequently exploited the vulnerabilities and defense mechanisms embedded within these LLMs.

Titled "MASTERKEY: Automated Jailbreaking of Large Language Model Chatbots", an updated and revised version of the study was published in October of 2023 to the arXiv server.

The study's objective was to explore and understand the defenses of LLM chatbots against "jailbreak" attacks. "Jailbreaking" is generally understood as any attempt to bypass the safeguards or restrictions built into a system. In the context of LLMs like ChatGPT, Bard, or Bing Chat, prompts are typically crafted to trick or exploit the model into performing actions or generating responses that it's programmed to avoid.

The general idea is to try and have the AI violate its content restrictions and have it circumvent its own filters and guidelines to generate responses that it normally wouldn't due to ethical, legal, or safety reasons. A common example is trying to have an LLM generate racist or hateful responses. Another common interest amongst jailbreaking enthusiasts is to craft prompts that lead the AI model to behave in unexpected ways, such as breaking out of its intended conversational scope or revealing internal mechanisms. Sometimes users will use insider knowledge about the model's

architecture or training data to create prompts that exploit specific weaknesses or oversights in the model's design. This is usually done to try to understand how the LLM works and to try to get a peek into its inner workings.

According to the study, traditional jailbreak attempts were largely ineffective against mainstream LLM chatbots, hinting at advanced, undisclosed defense strategies employed by their service providers. So the team decided to pit one AI against another.

Ph.D. student Mr. Liu Yi shows a database of successful jailbreaking prompts that compromised AI chatbots (Image: Nanyang Technological University).

Enter MASTERKEY, an end-to-end attack framework developed by the researchers. This framework employed a novel approach: reverse-engineering these defenses using time-based analysis and creating an AI capable of generating effective jailbreak prompts. The results were significant, as MASTERKEY demonstrated a higher success rate in jailbreaking these chatbots compared to existing techniques.

"The paper presents a novel approach for automatically generating jailbreak prompts against fortified LLM chatbots," explained NTU Ph.D. student Mr. Liu Yi, who co-authored the paper. "Training an

LLM 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."

MASTERKEY revealed the likely use of dynamic, real-time content moderation and keyword filtering as part of the chatbots' defense mechanisms. This insight is critical as it unveils the sophisticated nature of the safeguards in place, previously not fully understood by the public or even some AI researchers.

The study indicates that a nuanced process of reverse-engineering was utilized, where the researchers employed time-based analysis to understand how these chatbots responded to varying inputs. This analysis focused on identifying patterns such as response delays and variations in answers to similar prompts, which were indicative of the chatbots' underlying content moderation or filtering strategies. By closely examining these time-based characteristics, the team could infer the defense strategies that these advanced AI systems utilized to maintain ethical and safe interactions.

In parallel, the development of an AI model tailored to generate jailbreak prompts was the linchpin of the MASTERKEY framework. This process entailed training a specialized LLM, fine-tuned to create prompts that could effectively circumvent the defense mechanisms of other LLM chatbots. The training regimen included exposing it to a diverse array of prompts and corresponding responses. This exposure enabled the MASTERKEY model to learn and adapt to the patterns and limitations inherent in the target LLMs' defenses.

Subsequently, the efficacy of these jailbreak prompts was rigorously tested across various LLMs, leading to a cycle of continuous refinement and improvement. This process was key to enhancing MASTERKEY's proficiency in generating increasingly effective jailbreak prompts.

The team came up with some pretty innovative methods to bypass the built-in safeguards of a chatbot, focusing on crafting prompts that could subtly evade the chatbot's ethical restrictions, thereby inducing it to respond. AI developers typically implement keyword

censoring mechanisms that identify and block responses to inputs containing certain flagged words, indicative of potentially inappropriate or sensitive topics.

To navigate around these keyword censoring systems, the researchers adopted a clever tactic. They designed prompts interspersed with additional spaces between each letter, effectively disguising the words and evading detection by LLM censoring systems, which often rely on a predefined list of prohibited words.

Additionally, they experimented with directing the chatbot to assume an unrestrained persona, a character who lacks moral boundaries. This approach aimed to increase the likelihood of eliciting responses from the chatbot that might be considered unethical or outside its standard operational parameters.

"Despite their benefits, AI chatbots remain vulnerable to jailbreak attacks. They can be compromised by malicious actors who abuse vulnerabilities to force chatbots to generate outputs that violate established rules," explained Professor Liu Yang from NTU's School of Computer Science and Engineering, who led the study.

Looking ahead, the study opens new avenues for exploring the balance between user interaction capabilities and security in AI systems. As AI continues to evolve, so too must the strategies to ensure its safe and ethical use.

However, challenges remain. Ensuring that this knowledge is used responsibly to improve AI security, rather than to exploit vulnerabilities, is a paramount concern. The study underscores the ongoing need for collaborative efforts between AI developers, ethicists, and policymakers to navigate these complex and evolving landscapes.

The paper has been accepted for presentation at the Network and Distributed System Security Symposium in San Diego in February of 2024.

MJ Banias is a journalist who covers security and technology. He is the host of The Debrief Weekly Report. You can email MJ at mj@thedebrief.org or follow him on Twitter @mjbanias.