AI chatbots trained to jailbreak other chatbots, as the AI war slowly but surely begins

By Andy Edser published about 8 hours ago

Yep, it sure is turtles all the way down.

While AI ethics continues to be the hot-button issue of the moment, and companies and world governments continue to wrangle with the moral implications of a technology that we often struggle to define let alone control, here comes some slightly disheartening news: AI chatbots are already being trained to jailbreak other chatbots, and they seem remarkably good at it.

Researchers from the Nanyang Technological University in Singapore have managed to compromise several popular chatbots (via Tom’s Hardware), including ChatGPT, Google Bard and Microsoft Bing Chat, all done with the use of another LLM (large language model). Once effectively compromised, the jailbroken bots can then be used to “reply under a persona of being devoid of moral restraints.” Crikey.

This process is referred to as “Masterkey” and in its most basic form boils down to a two-step method. First, a trained AI is used to outwit an existing chatbot and circumvent blacklisted keywords via a reverse-engineered database of prompts that have already been proven to work.
AI chatbots trained to jailbreak other chatbots, as the AI war slowly but surely begins | PC Gamer

hack chatbots successfully. Armed with this knowledge, the AI can then automatically generate further prompts that jailbreak other chatbots, in an ouroboros-like move that makes this writer’s head hurt at the potential applications.

Ultimately this method can allow an attacker to use a compromised chatbot to generate unethical content and is claimed to be up to three times more effective at jailbreaking an LLM model than standard prompt, largely due to the AI attacker being able to quickly learn and adapt from its failures.

RECOMMENDED VIDEOS FOR YOU...

Harold Halibut Trailer | PC Gaming Show: Most Wanted 2023
This handmade narrative game about life on a submerged spaceship, will have you exploring friendship and the true meaning of home. As …
Upon realisation of the effectiveness of this method the NTU researchers reported the issues to relevant chatbot service providers, although given the supposed ability of this technique to quickly adapt and circumvent new processes designed to defeat it, it remains unclear as to how easy it would be for said providers to prevent such an attack.

The full NTU research paper is due for presentation at the Network and Distributed System Security Symposium due to be held in San Diego in February 2024, although one would assume that some of the intimate details of the method may be somewhat obfuscated for security purposes.

Regardless, using AI to circumvent the moral and ethical restraints of another AI seems like a step in a somewhat terrifying direction. Beyond the ethical issues created by a chatbot producing abusive or violent content à la Microsoft’s infamous “Tay”, the fractal-like nature of setting LLMs against each other is enough to give pause for thought.

While as a species we seem to be rushing headlong into an AI future we sometimes struggle to understand, the potential for the technology to be used against itself for malicious purposes seems an ever-growing threat, and it remains to be seen if service providers and LLM creators can react swiftly enough to head off these concerns before they cause serious issue or harm.
Andy built his first gaming PC at the tender age of 12, when IDE cables were a thing and high resolution wasn’t. After spending over 15 years in the production industry overseeing a variety of live and recorded projects, he started writing his own PC hardware blog for a year in the hope that people might send him things. Sometimes they did.

Now working as a hardware writer for PC Gamer, Andy can be found quietly muttering to himself and drawing diagrams with his hands in thin air. It’s best to leave him to it.