# The chatbot trained to “cheat” ChatGPT and Bard. It tricks them into producing illegal and dangerous content

TECHNOLOGY **Carrol World** 4 days ago **REPORT**

A chatbot specially trained to bypass the systems that chatbot manufacturers – such as OpenAI (ChatGPT) and Google (Bard) – have provided to prevent potentially dangerous content from being produced.

It was called Masterkey and was developed by Nanyang Technological University (NTU) in Singapore, which decided to test to what extent it was possible to exploit a chatbot system, also based on a large language model, to “rip”, in short, systems like ChatGPT and Bard or even Copilot (formerly Bing Chat) from Microsoft.

These chatbots, in fact, include security systems that prevent the production of illegal or dangerous content, for example the recipe for creating an explosive device at home.

These systems are based on various methods: in some cases certain keywords are detected which indicate to the chatbot in question not to respond to the user’s textual request (the prompt).

“*These chatbots are vulnerable to jailbreak-type attacks* (i.e. aimed at activating features that the developer on duty purposely prevents, ed.)*where a user with malicious intent manipulates the prompt to reveal sensitive, proprietary, or malicious information contrary to usage policies*“, wrote the researchers in the paper published on Arxiv, underlining that the various attempts already made previously “were not effective” on the most popular chatbots.

First of all, NTU researchers carried out a reverse engineering operation – that is, starting from the final product to understand how it works – so as to understand the defenses implemented by chatbot producers to avoid content of a certain type.

Then the researchers created a method to automatically generate prompts that overcome the chatbots' defense mechanisms; and the method is continuously improved because from time to time, based on the success or otherwise of the attempt, it learns new methods.

"*Our approach generates attack prompts that achieve an average success rate of 21.58%, significantly exceeding the 7.33% success rate achieved by existing prompts*", the researchers reported.

One method, for example, was to add a space after each character to fool keyword detection systems. Thanks to these systems, researchers, for example, **they managed to have content produced relating to what steps to follow to attack a website's computer system**.

The researchers have already informed the chatbot manufacturers of the ways they have used to overcome their systems' defenses.

"*Masterkey* – the researchers continue – *paves the way for a new strategy for exposing vulnerabilities in large language models and reinforces the need for more robust defenses against such breaches*".