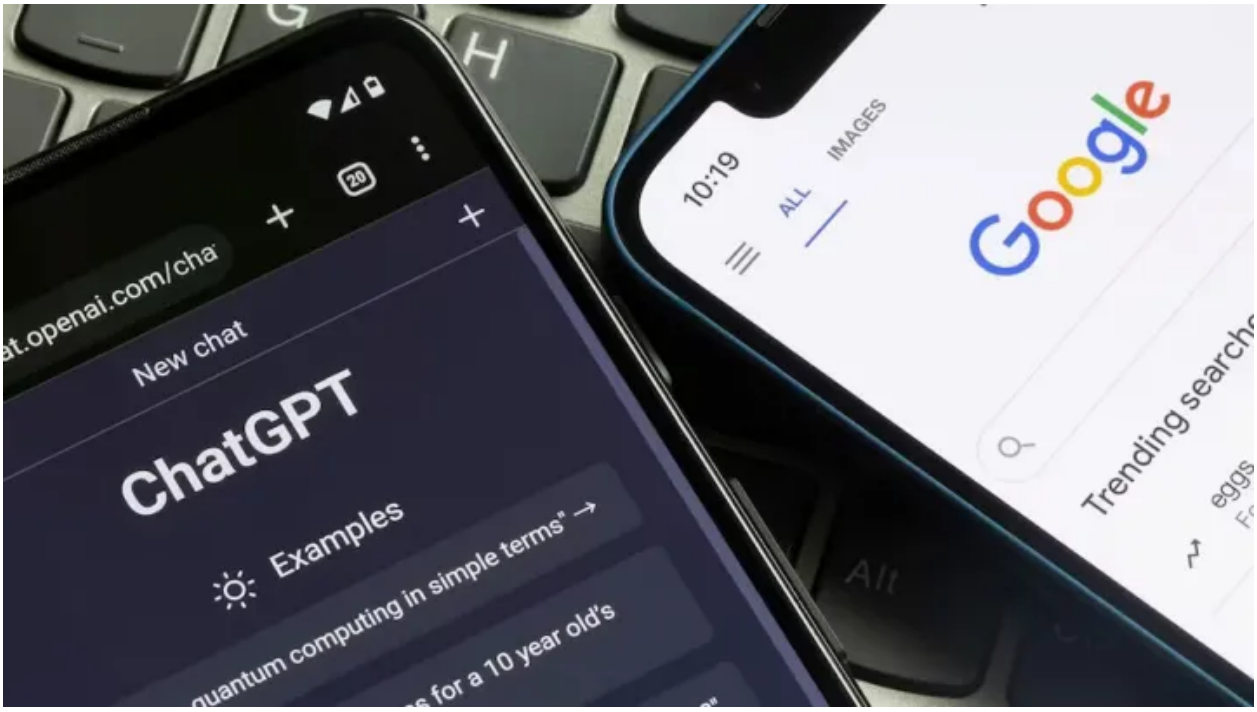
Researchers just unlocked ChatGPT

By Fionna Agomuoh

January 4, 2024 5:47AM

Researchers have discovered that it is possible to bypass the mechanism engrained in AI chatbots to make them able to respond to queries on banned or sensitive topics by using a different AI chatbot as a part of the training process.

A computer scientists team from Nanyang Technological University (NTU) of Singapore is unofficially calling the method a "jailbreak" but is more officially a "Masterkey" process. This system uses chatbots, including ChatGPT, Google Bard, and Microsoft Bing Chat, against one another in a two-part training method that allows two chatbots to learn each other's models and divert any commands against banned topics.



DigitalTrends

The team includes Professor Liu Yang and NTU Ph.D. students Mr. Deng Gelei and Mr. Liu Yi, who co-authored the research and developed the proof-of-concept attack methods, which essentially work like a bad actor hack.

Recommended Videos

According to the team, they first reverse-engineered one large language model (LLM) to expose its defense mechanisms. These would originally be blocks on the model and would not allow answers to certain prompts or words to go

Here's why people are claiming GPT-4 just got way better

2023 was the year of AI. Here were the 9 moments that defined it

This app just got me excited for the future of AI on Macs

But with this information reverse-engineered, they can teach a different LLM how to create a bypass. With the bypass created, the second model will be able to express more freely, based on the reverse-engineered LLM of the first model. The team calls this process a "Masterkey" because it should work even if LLM chatbots are fortified with extra security or are patched in the future.

> *The Masterkey process claims to be three times better at jailbreaking chatbots than prompts.*

Professor Lui Yang noted that the crux of the process is that it showcases how easily LLM AI chatbots can learn and adapt. The team claims its Masterkey process has had three times more success at jailbreaking LLM chatbots than a traditional prompt process. Similarly, some experts argue that the recently proposed glitches that certain LLMs, such as GPT-4 have been experiencing are signs of it becoming more advanced, rather than dumber and lazier, as some critics have claimed.

Since AI chatbots became popular in late 2022 with the introduction of OpenAI's ChatGPT, there has been a heavy push toward ensuring various services are safe and welcoming for everyone to use. OpenAI has put safety warnings on its ChatGPT product during sign-up and sporadic updates, warning of unintentional slipups in language. Meanwhile, various chatbot spinoffs have been fine to allow swearing and offensive language to a point.

Additionally, actual bad actors quickly began to take advantage of the demand for ChatGPT, Google Bard, and other chatbots before they became wildly available. Many campaigns advertised the products on social media with malware attached to image links, among other attacks. This showed quickly that AI was the next frontier of cybercrime.

The NTU research team contacted the AI chatbot service providers involved in the study about its proof-of-concept data, showing that jailbreaking for chatbots is real. The team will also present their findings at the Network and Distributed System Security Symposium in San Diego in February.

Editors' Recommendations

- OpenAI and Microsoft sued by NY Times for copyright infringement
- What is Grok? Elon Musk's controversial ChatGPT competitor explained
- Google might finally have an answer to Chat GPT-4
- One year ago, ChatGPT started a revolution
- Here's why people are saying GPT-4 is getting 'lazy'

Topics

Artificial Intelligence Tech News

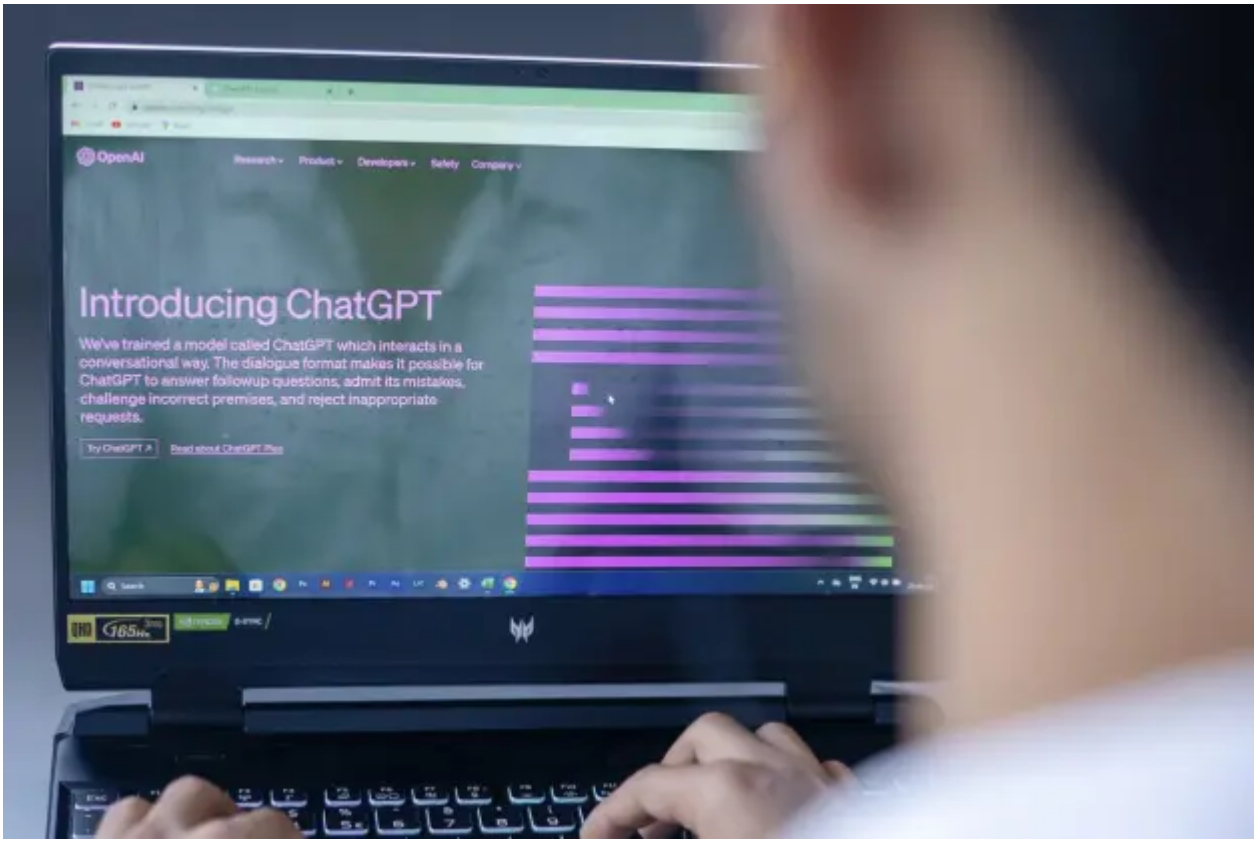
Fionna Agomuoh

Computing Writer

Fionna Agomuoh is a technology journalist with over a decade of experience writing about various consumer electronics topics...

COMPUTING

Here's why you can't sign up for ChatGPT Plus right now



Read more

COMPUTING

OpenAI is on fire – here's what that means for ChatGPT and Windows

Read more

COMPUTING

The world responds to the creator of ChatGPT being fired by his own company

Read more