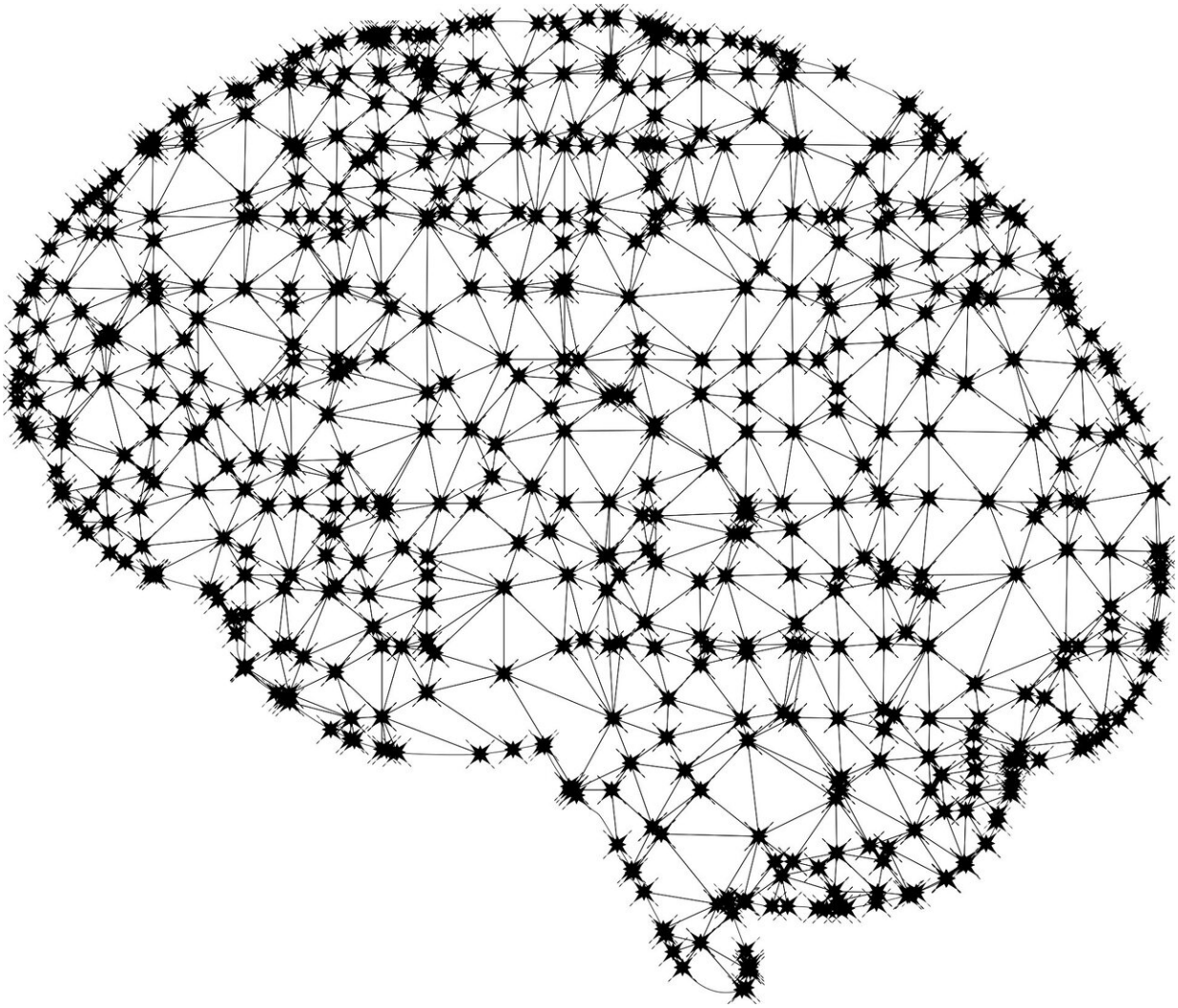
AI researchers expose critical vulnerabilities within major large language models

October 12 2023



Credit: CC0 Public Domain

Large Language Models (LLMs) such as ChatGPT and Bard have taken the world by storm this year, with companies investing millions to develop these AI tools, and some leading AI chatbots being valued in the billions.

These LLMs, which are increasingly used within AI chatbots, scrape the entire Internet of information to learn and to inform answers that they provide to user-specified requests, known as "prompts."

However, computer scientists from the AI security start-up Mindgard and Lancaster University in the UK have demonstrated that chunks of these LLMs can be copied in less than a week for as little as $50, and the information gained can be used to launch targeted attacks.

The researchers warn that attackers exploiting these vulnerabilities could reveal private confidential information, bypass guardrails, provide incorrect answers, or stage further targeted attacks.

Detailed in a new paper to be presented at CAMLIS 2023 (Conference on Applied Machine Learning for Information Security) the researchers show that it is possible to copy important aspects of existing LLMs cheaply, and they demonstrate evidence of vulnerabilities being transferred between different models.

This attack, termed "model leeching," works by talking to LLMs in such a way—asking it a set of targeted prompts—so that the LLMs elicit insightful information giving away how the model works.

The research team, which focused their study on ChatGPT-3.5-Turbo, then used this knowledge to create their own copy model, which was 100 times smaller but replicated key aspects of the LLM.

The researchers were then able to use this model copy as a testing

ground to work out how to exploit vulnerabilities in ChatGPT without detection. They were then able to use the knowledge gleaned from their model to attack vulnerabilities in ChatGPT with an 11% increased success rate.

Dr. Peter Garraghan of Lancaster University, CEO of Mindgard, and Principal Investigator on the research, said, "What we discovered is scientifically fascinating, but extremely worrying. This is among the very first works to empirically demonstrate that security vulnerabilities can be successfully transferred between closed source and open source Machine Learning models, which is extremely concerning given how much industry relies on publicly available Machine Learning models hosted in places such as HuggingFace."

The researchers say their work highlights that although these powerful digital AI technologies have clear uses, there exist hidden weaknesses, and there may even be common vulnerabilities across models.

Businesses across industry are currently or preparing to invest billions in creating their own LLMs to undertake a wide range of tasks such as smart assistants. Financial services and large enterprises are adopting these technologies but researchers say that these vulnerabilities should be a major concern for all businesses that are planning to build or use third party LLMs.

Dr. Garraghan said, "While LLM technology is potentially transformative, businesses and scientists alike will have to think very carefully on understanding and measuring the cyber risks associated with adopting and deploying LLMs."

Provided by Lancaster University