UK cybersecurity agency warns of chatbot ‘prompt injection’ attacks

Publication Date: 2023-08-29

Author: Hibaq Farah

Section: Technology

Tags: Chatbots, Hacking, Artificial intelligence (AI), ChatGPT, news

Article URL: https://www.theguardian.com/technology/2023/aug/30/uk-cybersecurity-agency-warns-of-chatbot-prompt-injection-attacks

The UK’s cybersecurity agency has warned that chatbots can be manipulated by hackers to cause scary real-world consequences. The National Cyber Security Centre (NCSC) has said there are growing cybersecurity risks of individuals manipulating the prompts through “prompt injection” attacks. This is where a user creates an input or a prompt that is designed to make a language model – the technology behind chatbots – behave in an unintended manner. A chatbot runs on artificial intelligence and is able to give answers to prompted questions by users. They mimic human-like conversations, which they have been trained to do through scraping large amounts of data. Commonly used in online banking or online shopping, chatbots are generally designed to handle simple requests. Large language models (LLMs), such as OpenAI’s ChatGPT and Google’s AI chatbot Bard, are trained using data that generates human-like responses to user prompts. Since chatbots are used to pass data to third-party applications and services, the NCSC has said that risks from malicious prompt injection will grow. For instance, if a user inputs a statement or question that a language model is not familiar with, or if they find a combination of words to override the model’s original script or prompts, the user can cause the model to perform unintended actions. Such inputs could cause a chatbot to generate offensive content or reveal confidential information in a system that accepts unchecked input. This year, Microsoft released a new version of its Bing search engine and conversational bot powered by LLMs. A Stanford university student, Kevin Liu, was able to create a prompt injection to find Bing Chat’s initial prompt. The entire prompt of Microsoft’s Bing Chat, a list of statements written by Open AI or Microsoft that determine how the chatbot interacts with users, which is hidden from users, was revealed by Liu putting in a prompt that requested the Bing Chat “ignore previous instructions”. The security researcher Johann Rehberger found that he could force ChatGPT to respond to new prompts through a third party that he did not initially request. Rehberger ran a prompt injection through YouTube transcripts and found that ChatGPT could access YouTube transcripts, which could cause more indirect prompt injection vulnerabilities. According to the NCSC, prompt injection attacks can also cause real-world consequences if systems are not designed with security. The vulnerability of chatbots and the ease with which prompts can be manipulated could cause attacks, scams and data theft. LLMs are increasingly used to pass data to third-party applications and services, meaning the risks from malicious prompt injection will grow. The NCSC said: “Prompt injection and data poisoning attacks can be extremely difficult to detect and mitigate. “However, no model exists in isolation, so what we can do is design the whole system with security in mind. That is, by being aware of the risks associated with the ML [machine learning] component, we can design the system in such a way as to prevent exploitation of vulnerabilities leading to catastrophic failure. “A simple example would be applying a rules-based system on top of the ML model to prevent it from taking damaging actions, even when prompted to do so.” The NCSC says that cyber-attacks caused by artificial intelligence and machine learning that leaves systems vulnerable can be mitigated through designing for security and understanding the attack techniques that exploit “inherent vulnerabilities” in machine learning algorithms.