Microsoft and OpenAI say hackers are using ChatGPT to improve cyberattacks

Microsoft and OpenAI are revealing today that hackers are already using large language models like ChatGPT to refine and improve their existing cyberattacks. In newly published research, Microsoft and OpenAI have detected attempts by Russian, North Korean, Iranian, and Chinese-backed groups using tools like ChatGPT for research into targets, to improve scripts, and to help build social engineering techniques.

“Cybercrime groups, nation-state threat actors, and other adversaries are exploring and testing different AI technologies as they emerge, in an attempt to understand potential value to their operations and the security controls they may need to circumvent,” says Microsoft in a blog post today.

The Strontium group, linked to Russian military intelligence, has been found to be using LLMs “to understand satellite communication protocols, radar imaging technologies, and specific technical parameters.” The hacking group, known also as APT28 or Fancy Bear, has been active during the Russia-Ukraine war and was previously involved in targeting Hillary Clinton’s presidential campaign in 2016.

The group has also been using LLMs to help with “basic scripting tasks, including file manipulation, data selection, regular expressions, and multiprocessing, to potentially automate or optimize technical operations,” according to Microsoft.

A North Korean hacking group, known as Thallium, has been using LLMs to research publicly reported vulnerabilities and target organizations, to aid in basic scripting tasks, and to draft content for phishing campaigns. Microsoft says the Iranian group known as Curium has also been using LLMs to generate phishing emails and even code for avoiding detection by antivirus applications. Chinese state-affiliated hackers are also using LLMs for research, scripting, translations, and to refine their existing tools.

There have been fears about the use of AI in cyberattacks, particularly as AI tools like WormGPT and FraudGPT have emerged to assist in the creation of malicious emails and cracking tools. A senior official at the National Security Agency also warned last month that hackers are using AI to make their phishing emails look more convincing.

Microsoft and OpenAI haven't detected any "significant attacks" using LLMs yet, but the companies have been shutting down all accounts and assets associated with these hacking groups. "At the same time, we feel this is important research to publish to expose early-stage, incremental moves that we observe well-known threat actors attempting, and share information on how we are blocking and countering them with the defender community," says Microsoft.

While the use of AI in cyberattacks appears to be limited right now, Microsoft does warn of future use cases like voice impersonation. "AI-powered fraud is another critical concern. Voice synthesis is an example of this, where a three-second voice sample can train a model to sound like anyone," says Microsoft. "Even something as innocuous as your voicemail greeting can be used to get a sufficient sampling."

Naturally, Microsoft's solution is using AI to respond to AI attacks. "AI can help attackers bring more sophistication to their attacks, and they have resources to throw at it," says Homa Hayatyfar, principal detection analytics manager at

Microsoft. “We’ve seen this with the 300+ threat actors Microsoft tracks, and we use AI to protect, detect, and respond.”

Microsoft is building a Security Copilot, a new AI assistant that’s designed for cybersecurity professionals to identify breaches and better understand the huge amount of signals and data that’s generated through cybersecurity tools daily. The software giant is also overhauling its software security following major Azure cloud attacks and even Russian hackers spying on Microsoft executives.

https://www.theverge.com/2024/2/14/24072706/microsoft-openai-cyberattack-tools-ai-chatgpt