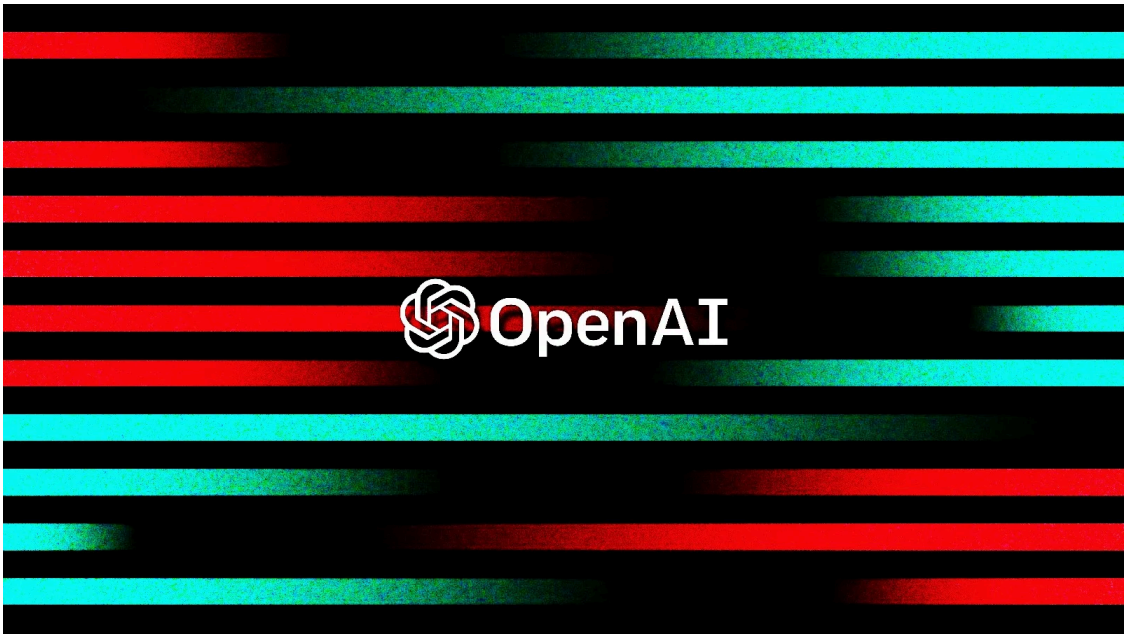
OpenAI bans ChatGPT accounts used by North Korean hackers

By Sergiu Gatlan

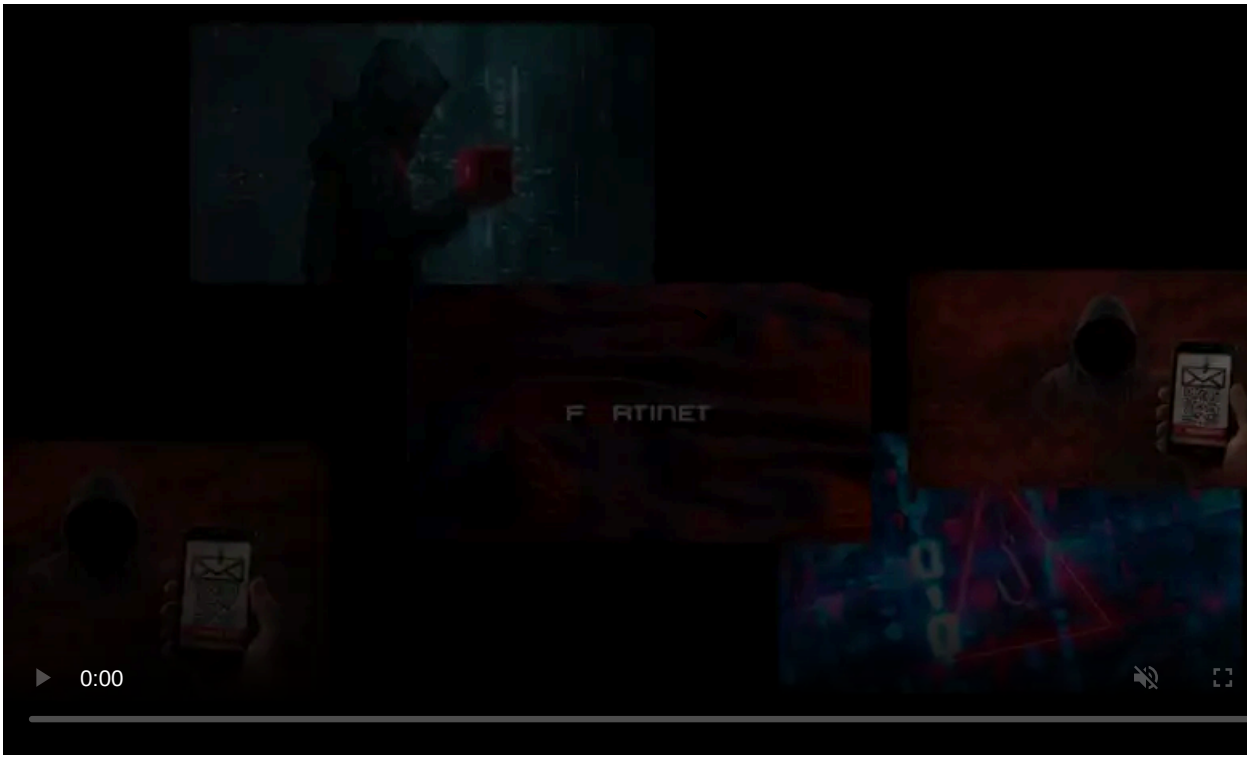
Published: 2025-02-24 · Archived: 2026-04-05 20:37:16 UTC

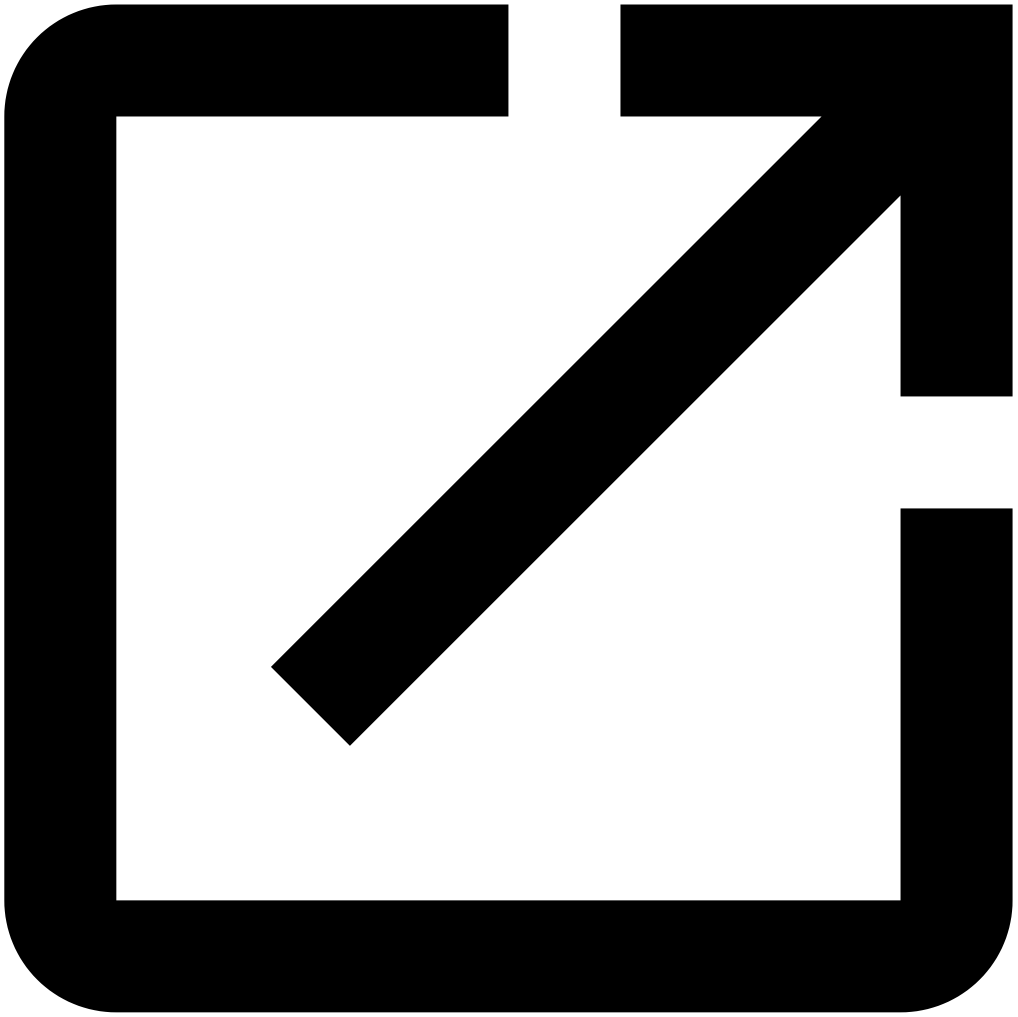


OpenAI says it blocked several North Korean hacking groups from using its ChatGPT platform to research future targets and find ways to hack into their networks.

"We banned accounts demonstrating activity potentially associated with publicly reported Democratic People's Republic of Korea (DPRK)-affiliated threat actors," the company said in its February 2025 threat intelligence report.

"Some of these accounts engaged in activity involving TTPs consistent with a threat group known as VELVET CHOLLIMA (AKA Kimsuky, Emerald Sleet), while other accounts were potentially related to an actor that was assessed by a credible source to be linked to STARDUST CHOLLIMA (AKA APT38, Sapphire Sleet)."


F RTINET
0:00



The now-banned accounts were detected using information from an industry partner. In addition to researching what tools to use during cyberattacks, the threat actors used ChatGPT to find information on cryptocurrency-related topics, which are common interests linked to North Korean state-sponsored threat groups.

The malicious actors also used ChatGPT for coding assistance, including help on how to use open-source Remote Administration Tools (RAT), as well as debugging, researching, and development assistance for open-source and publicly available security tools and code that could be used in Remote Desktop Protocol (RDP) brute force attacks.

OpenAI threat analysts also found that the North Korean actors revealed staging URLs for malicious binaries unknown to security vendors at the time while debugging auto-start extensibility point (ASEP) locations and macOS attack techniques.

These staging URLs and the associated compiled executable files were submitted to an online scanning service to facilitate sharing with the broader security community. As a result, some vendors now reliably detect these binaries, protecting potential victims from future attacks.

Other malicious activity uncovered by OpenAI while researching in what ways the North Korean threat actors used the banned accounts includes but is not limited to:

- Asking about vulnerabilities in various applications,
- Developing and troubleshooting a C#-based RDP client to enable,
- Requesting code to bypass security warnings for unauthorized RDP,
- Requested numerous PowerShell scripts for RDP connections, file upload/download, executing code from memory, and obfuscating HTML content,
- Discusses creating and deploying obfuscated payloads for execution,
- Seeking methods to conduct targeted phishing and social engineering against cryptocurrency investors and traders, as well as more generic phishing content,
- Crafting phishing emails and notifications to manipulate users into revealing sensitive information.

The company also banned accounts linked to a potential North Korean IT worker scheme, described as having all the characteristics of efforts to obtain income for the Pyongyang regime by tricking Western companies into hiring North Koreans.

"After appearing to gain employment they used our models to perform job-related tasks like writing code, troubleshooting and messaging with coworkers," OpenAI explained. "They also used our models to devise cover stories to explain unusual behaviors such as avoiding video calls, accessing corporate systems from unauthorized countries or working irregular hours."

Since October 2024, when it published its previous report, OpenAI has also detected and disrupted two campaigns originating from China, "Peer Review" and "Sponsored Discontent." These campaigns used the ChatGPT models to research and develop tools linked to a surveillance operation and generate anti-American, Spanish-language articles.

In the October report, OpenAI revealed that since the beginning of 2024, it disrupted over twenty campaigns linked to cyber operations and covert influence operations associated with Iranian and Chinese state-sponsored hackers.
