ChatGPT at Risk: The Latest AI Package Hallucination Cyberattack

ChatGPT is at risk of a new cyberattack. According to recent research conducted by Vulcan Cyber, hackers can exploit the chatbot to disseminate malicious packages within the developer's group. This attack, known as "AI Package Hallucination," involves the creation of deceptive URLs, references, or complete code libraries and functions that do not actually exist.

Using this new technique, cybercriminals have the ability to substitute unpublished packages with their own malicious counterparts. This enables attackers to carry out supply chain attacks, incorporating malevolent libraries into well-known storage systems.

Let's delve deeper into this matter to fully understand the gravity of the situation.

What Is ChatGPT?

ChatGPT is a generative AI chatbot that uses the natural language processing (NLP) method to create humanlike conversational dialogue. The language model can answer questions and help users with various tasks, such as writing essays, composing songs, creating social media posts, and developing codes.

What Is AI Package Hallucination?

"AI Package Hallucination" is the latest, one of the most critical and deadly hacking attacks that ChatGPt has faced to date. Using this technique, the cyber offenders can now transfer malicious packages straight to the developer's team.

Recently, the researchers at Vulcan Cyber have identified a concerning trend. This attack vector involves the manipulation of web URLs, references, and complete code libraries and functions that simply do not exist.

Vulcan's analysis has revealed that this anomaly may be attributed to the utilization of outdated training data in ChatGPT, resulting in the recommendation of non-existent code libraries.

The researchers have issued a warning regarding the potential exploitation of this vulnerability. They caution that hackers can gather the names of these non-existent packages and create their own malicious versions. Subsequently, unsuspecting developers may inadvertently download these malicious packages based on the recommendations provided by ChatGPT.

This underscores the urgent need for vigilance within the developer community to prevent unwittingly incorporating harmful code into their projects.

What Do The Researchers Say?

Vulcan researchers evaluated ChatGPT by testing it with common questions sourced from the Stack Overflow coding platform. They specifically inquired about these questions within the Python and Node.js environments to assess ChatGPT's capabilities in these programming languages.

The researchers extensively queried ChatGPT with over 400 questions, and during this evaluation, approximately 100 of its answers included at least one reference to Python or Node.js packages that do not exist in reality.

As a result, ChatGPT's responses involved mentioning 150 non-existent packages in total.

The researchers highlighted a potential security concern regarding the use of ChatGPT's package recommendations. They expressed that attackers could take advantage of the suggested package names by ChatGPT, creating their own malicious versions and uploading them to popular software repositories. Consequently, developers who rely on ChatGPT for coding solutions may unknowingly download and install these malicious packages.

The researchers emphasized that the impact of such a scenario would be significantly more dangerous if developers searching for coding solutions online ask ChatGPT for package recommendations and inadvertently end up utilizing a malicious package.

How Does AI Package Hallucination Work?

The vice president of security operations at Ontinue, Craig Jones, has told how the AI Package Hallucination attack could work:

- Attackers ask ChatGPT for coding help in common tasks;
- ChatGPT might provide a package recommendation that either doesn't exist or is unpublished yet (a "hallucination");
- Then, the attackers create a malicious version of that recommended package and publish it;
- As such, while other developers query the same questions to ChatGPT, it might recommend the same currently existing but malicious package to them.

Precautionary Steps To Prevent the Attack

Melissa Bischoping, director of endpoint security research at Tanium, emphasizes the importance of cautious code execution practices in light of the recent cyberattack:

You should never download and execute code you don't understand and haven't tested by just grabbing it from a random source – such as open-source GitHub repos or ChatGPT's recommendations.

Additionally, Bischoping recommends maintaining private copies of code rather than importing directly from public repositories, as these have been compromised in the ongoing cyberattack.

Use of this strategy will continue, and the best defense is to employ secure coding practices and thoroughly test and review code intended for use in production environments.

According to Vulcan Cyber, there are several precautionary steps that developers can take to identify potentially malicious packages and protect themselves from cyberattacks. These steps include:

1. Checking the package creation date: If a package was recently created, it might raise suspicions.
2. Evaluating the number of downloads: If a package has very few or no downloads, it may be less reliable and should be approached with caution.
3. Reviewing comments and ratings: If a package has a lack of comments or stars, it may be prudent to exercise caution before installing it.
4. Examining attached notes or documentation: If accompanying documentation or notes associated with the package are incomplete, misleading, or raise suspicions, it is advisable to think twice before proceeding with the installation.

By remaining vigilant and following these precautionary steps, developers can minimize the risk of falling victim to a cyberattack through ChatGPT or any other code execution environment.

The Bottom Line

The Vulcan Cyber research team's discovery of the AI Hallucination Attack on the chatbot highlights the significant threat it poses to users relying on ChatGPT for their daily work.

In order to protect themselves from this attack and the potential risks associated with malicious packages, developers and other potential victims should exercise extreme caution and adhere to primary security guidance.