INSIGHTS

Cyber made simple.

AI CONTINUES TO HAVE PROBLEMS

BY JAY BORDEN

AI fever is still raging as more companies enter the AI market place and more companies look for ways to use it.

Google's AI offering is called Gemini, formerly called Bard, and is supposed to have protections in place to prevent misuse or delivering harmful content.

The researchers found Gemini Advanced, the API version, appears to provide access that circumvents the protections. The result: misinformation, details on how to steal a car, and more.

Using the APIs the model itself can be manipulated to provide system prompts. This is especially troubling as it allows access to where the basic instructions and guidelines are defined in the system. Working at this level allows any limits to be removed or redefined to suit anyone.

Reaching the system prompts was not difficult. The researchers could trick Gemini into revealing the secret passphrase that protects the system prompt. Two tries and…success!

This level of access let the researchers get Gemini to create a programming shell on the server itself. Using that allowed them to find and extract any information on the system, including confidential data input by other users of Gemini.

If a Gemini user in any company entered confidential information for it to prepare an answer, report, presentation, or anything, attackers will have it.

The researchers asked Gemini to write a fictional story about the upcoming election and Gemini invoked its protections and refused. However, all they needed to do was instruct Gemini to enter a fictional state and write the story and it did.

onebrightlycyber.com
(888) 773-1920
A global leader in cyber service, technology, insurance and innovation.

AI CONTINUES TO HAVE PROBLEMS

Continued

This could be distributed as misinformation about the election. Or any other topic attackers want to manipulate.

Gemini, similar to other AI systems could be easily manipulated into providing nonsense answers, false answers, or even the secret keys.

So much for privacy and confidentiality and accuracy.

On another note, infostealers, malware that infects a device and steals information then sends it to cybercriminals, are on the rise. This trend is viewed as a cause of ChatGPT credential leaks increasing 36% between the first 4 months and the last 4 months of 2023.

At least 225,000 sets of ChatGPT credentials are for sale on the Darkweb. Any cybercriminal can purchase the credentials and using them can log into ChatGPT as a legitimate user and access anything that user can. No manipulation of the system needed.

That would include any confidential information that a user or a colleague may have entered into the system for ChatGPT to employ.

ChatGPT is used by developers to optimize application code and attackers with stolen credentials would get that code. It could then be used to demand payment from the company to prevent public release or the code can be altered to provide backdoors into the system.

More AI news. Elon Musk initially invested in OpenAI, the makers of ChatGPT, but after a disagreement divested his holdings. Now he is developing his own AI system and said it will be open source and offered for free. The effect and risks of that on the AI market are yet to be seen.

Amazon has a chatbot to help people find relevant reviews of products on the site. However, it is not limited to that. Researchers discovered it to be little different from other AI offerings in that it was found to recommend racist books, write resumes based upon fictitious work experience it created, and provide completely false information about the work conditions at Amazon.

AI CONTINUES TO HAVE PROBLEMS

Continued

The app is called “Ask about his product” and can be found on the Amazon mobile app.

Without a doubt AI and Large Learning models provide benefits. But they do come with risks. Companies need to determine how these systems can be used safely. Polices on use should be issued and enforced.

Stay tuned as this is continuing to unfold.

To learn all the ways we can help make your company and family safer, visit onebrightlycyber.com, contact OneBrightlyCyber at info@onebrightlycyber.com, or call (888) 773-1920.