The Information

Watch Out for Model Hijacking

Legit Security Uncovers Supply Chain Weakness in Popular Hugging Face Repository

By Stephanie Palazzolo

AI AGENDA AI

Dec. 7, 2023 7:15 AM PST

Over the past year, we've seen plenty of examples of large langu age models spitting out worrisome content, from encouraging users to harm themselves to providing them with instructions on how to build bombs. (Just yesterday, we featured a new security threat to LLMs.) These incidents have caught the eye of security experts and raised concerns around the safety of generative AI models.

As headline-grabbing as these examples are, though, there's a much less sexy but equally dangerous security threat in AI: supply chain attacks. In plain English, supply chain attacks refer to a type of cyber-attack that targets a vendor who sells services or software used by other companies to build their products.

Cybersecurity startup Legit Security recently uncovered a supply chain weakness in the popular open-source model repository Hugging Face, highlighting some of the often-overlooked downsides of using open-source AI.

Commonly, developers will use open-source models or datasets from the Hugging Face platform when building AI-powered products. When the owners of these models or datasets update or rename them, any products using the models or datasets automatically begin using the updated versions to prevent disruptions, said Liav Caspi, co-founder of Legit Security.

The problem arises when a developer registers a new model or dataset under the old name. When that happens, products can start accidentally using that model or dataset instead, Caspi explained. That allows attackers to create new malicious models that borrow names from legitimate ones in order to feed poisoned data or malware into users' apps, he said. Legit Security used the hack to prove that hundreds of thousands of developer apps that relied on the "ai-forever" models could quickly be corrupted, he said.

Legit Security dubbed this bug "AIJacking." And when Caspi's team alerted Hugging Face to the bug, Hugging Face said it had a mechanism that automatically retired old names of the most popular models and datasets once developers renamed them. That leaves Hugging Face to manually retire less-popular models, which Caspi worried isn't a scalable solution. A spokesperson for Hugging Face declined to comment.

The incident shows that while regulators and doomers worry about AI ending humanity, or terrorists creating bioweapons with open-source models, a variety of more immediate security threats looms right now.

Developers can guard against AIJacking attacks by referencing a specific version of a model hosted by Hugging Face, and preventing their apps from automatically executing code from third-party vendors like Hugging Face, Caspi said.

The AIJacking flaw is also a stark reminder of the dangers of building software on top of young and relatively untested startups, as we saw with the controversy surrounding OpenAI in recent weeks. That debacle prompted a number of customers to consider switching to OpenAI models offered by a tech veteran like Microsoft instead.