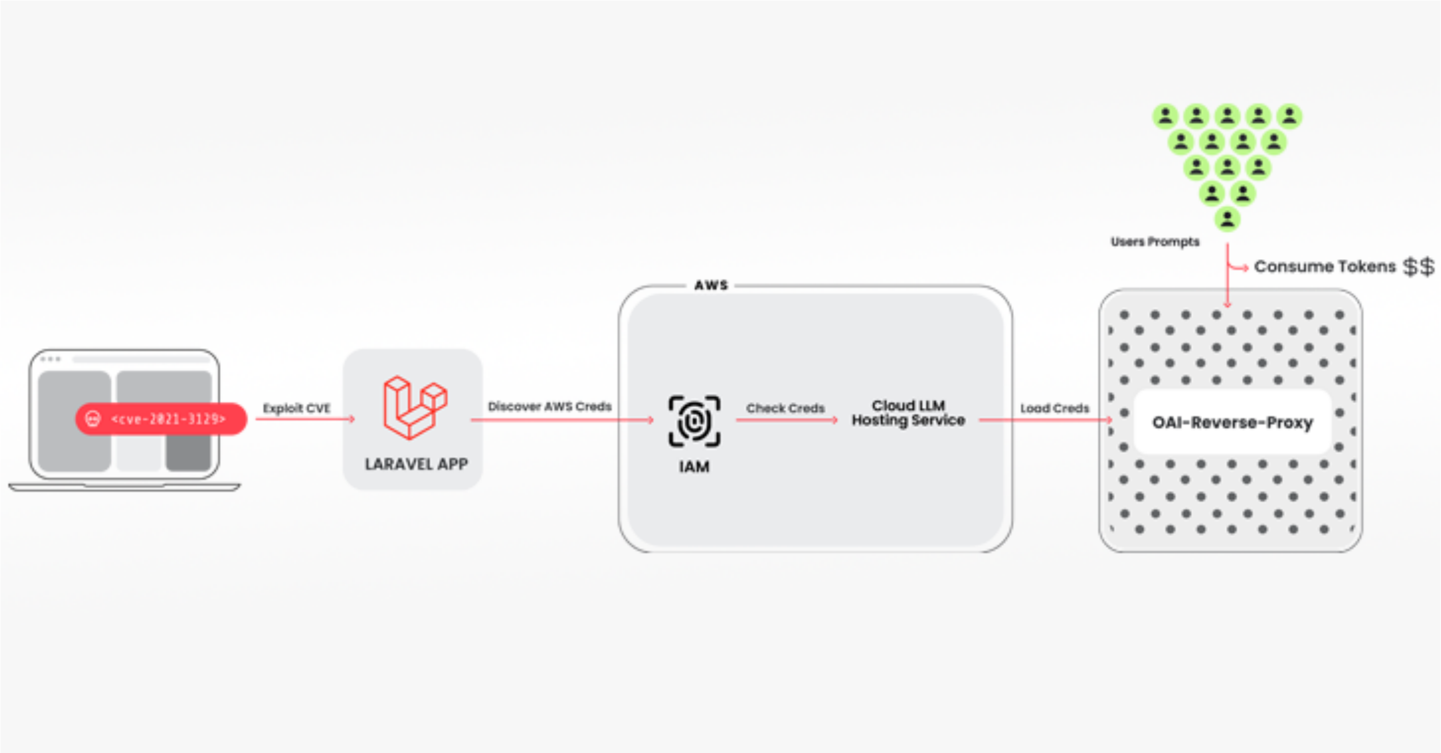
Researchers Uncover 'LLMjacking' Scheme Targeting Cloud-Hosted AI Models

May 10, 2024 Newsroom



Cybersecurity researchers have discovered a novel attack that employs stolen cloud credentials to target cloud-hosted large language model (LLM) services with the goal of selling access to other threat actors.

The attack technique has been codenamed **LLMjacking** by the Sysdig Threat Research Team.

"Once initial access was obtained, they exfiltrated cloud credentials and gained access to the cloud environment, where they attempted to access local LLM models hosted by cloud providers," security researcher Alessandro Brucato said. "In this instance, a local Claude (v2/v3) LLM model from Anthropic was targeted."

The intrusion pathway used to pull off the scheme entails breaching a system running a vulnerable version of the Laravel Framework (e.g., CVE-2021-3129), followed by getting hold of Amazon Web Services (AWS) credentials to access the LLM services.

Among the tools used is an open-source Python script that checks and validates keys for various offerings from Anthropic, AWS Bedrock, Google Cloud Vertex AI, Mistral, and OpenAI, among others.

"No legitimate LLM queries were actually run during the verification phase," Brucato explained. "Instead, just enough was done to figure out what the credentials were capable of and any quotas."

The keychecker also has integration with another open-source tool called oai-reverse-proxy that functions as a reverse proxy server for LLM APIs, indicating that the threat actors are likely providing access to the compromised accounts without actually exposing the underlying credentials.

"If the attackers were gathering an inventory of useful credentials and wanted to sell access to the available LLM models, a reverse proxy like this could allow them to monetize their efforts," Brucato said.

Furthermore, the attackers have been observed querying logging settings in a likely attempt to sidestep detection when using the compromised credentials to run their prompts.

The development is a departure from attacks that focus on prompt injections and model poisoning, instead allowing attackers to monetize their access to the LLMs while the owner of the cloud account foots the bill without their knowledge or consent.

Sysdig said that an attack of this kind could rack up over $46,000 in LLM consumption costs per day for the victim.

"The use of LLM services can be expensive, depending on the model and the amount of tokens being fed to it," Brucato said. "By maximizing the quota limits, attackers can also block the compromised organization from using models legitimately, disrupting business operations."

Organizations are recommended to enable detailed logging and monitor cloud logs for suspicious or unauthorized activity, as well as ensure that effective vulnerability management processes are in place to prevent initial access.