AI Could Generate 10,000 Malware Variants, Evading Detection in 88% of Case

The Hacker News ⋮ 6-7 minutes

Cybersecurity researchers have found that it's possible to use large language models (LLMs) to generate new variants of malicious JavaScript code at scale in a manner that can better evade detection.

"Although LLMs struggle to create malware from scratch, criminals can easily use them to rewrite or obfuscate existing malware, making it harder to detect," Palo Alto Networks Unit 42 researchers said in a new analysis. "Criminals can prompt LLMs to perform transformations that are much more natural-looking, which makes detecting this malware more challenging."

With enough transformations over time, the approach could have the advantage of degrading the performance of malware classification systems, tricking them into believing that a piece of nefarious code is actually benign.

While LLM providers have increasingly enforced security guardrails to prevent them from going off the rails and producing unintended output, bad actors have advertised tools like WormGPT as a way to automate the process of crafting convincing phishing emails that are tailed to prospective targets and even create novel malware.

Back in October 2024, OpenAI disclosed it blocked over 20 operations and deceptive networks that attempt to use its platform for reconnaissance, vulnerability research, scripting support, and debugging.

Unit 42 said it harnessed the power of LLMs to iteratively rewrite existing malware samples with an aim to sidestep detection by machine learning (ML) models like Innocent Until Proven Guilty (IUPG) or PhishingJS, effectively paving the way for the creation of 10,000 novel JavaScript variants without altering the functionality.

The adversarial machine learning technique is designed to transform the malware using various methods -- namely, variable renaming, string splitting, junk code insertion, removal of unnecessary whitespaces, and a complete reimplementation of the code -- every time it's fed into the system as input.

"The final output is a new variant of the malicious JavaScript that maintains the same behavior of the original script, while almost always having a much lower malicious score," the company said, adding the greedy algorithm flipped its own malware classifier model's verdict from malicious to benign 88% of the time.

To make matters worse, such rewritten JavaScript artifacts also evade detection by other malware analyzers when uploaded to the VirusTotal platform.

Another crucial advantage that LLM-based obfuscation offers is that its lot of rewrites look a lot more natural than those achieved by libraries like obfuscator.io, the latter of which are easier to reliably detect and fingerprint owing to the manner they introduce changes to the source code.

"The scale of new malicious code variants could increase with the help of generative AI," Unit 42 said. "However, we can use the same tactics to rewrite malicious code to help generate training data that can improve the robustness of ML models."

TPUXtract Attack Targets Google Edge TPUs#

The disclosure comes as a group of academics from North Carolina State University devised a side-channel attack dubbed TPUXtract to conduct model stealing attacks on Google Edge Tensor Processing Units (TPUs) with 99.91% accuracy. This could then be exploited to facilitate intellectual property theft or follow-on cyber attacks.

"Specifically, we show a hyperparameter stealing attack that can extract all layer configurations including the layer type, number of nodes, kernel/filter sizes, number of filters, strides, padding, and activation function," the researchers said. "Most notably, our attack is the first comprehensive attack that can extract previously unseen models."

The black box attack, at its core, captures electromagnetic signals emanated by the TPU when neural network inferences are underway – a consequence of the computational intensity associated with running offline ML models – and exploits them to infer model hyperparameters. However, it hinges on the adversary having physical access to a target device, not to mention possessing expensive equipment to probe and obtain the traces.

"Because we stole the architecture and layer details, we were able to recreate the high-level features of the AI," Aydin Aysu, one of the authors of the study, said. "We then used that information to recreate the functional AI model, or a very close surrogate of that model."

EPSS Found Susceptible to Manipulation Attacks#

Last week, Morphisec also disclosed that AI frameworks like the Exploit Prediction Scoring System (EPSS), which is used by a wide range of security vendors, could be subjected to adversarial attacks, affecting how it evaluates risk and the likelihood of a known software vulnerability being exploited in the wild.

"The attack targeted two key features in EPSS's feature set: social media mentions and public code availability," security researcher Ido Ikar said, adding it's possible to influence the model's output by "artificially inflating these indicators" by sharing random posts on X about a security flaw and creating a GitHub repository containing an empty file that contains an exploit for it.

The proof-of-concept (PoC) technique shows that a threat actor could leverage EPSS' reliance on external signals to boost the activity metrics of specific CVEs, potentially "misguiding" organizations that count on the EPSS scores to prioritize their vulnerability management efforts.

"Following the injection of artificial activity through generated social media posts and the creation of a placeholder exploit repository, the model's predicted probability for exploitation increased from 0.1 to 0.14," Ikar noted. "Additionally, the percentile ranking of the vulnerability rose from the 41st percentile to the 51st percentile, pushing it above the median level of perceived threat."

Found this article interesting? Follow us on Twitter and LinkedIn to read more exclusive content we post.