MITRE: Russian APT28's LameHug, a Pilot for Future AI Cyber-Attacks

By Kevin Poireault

Published: 2025-08-12 · Archived: 2026-04-05 14:25:19 UTC

APT28's LameHug wasn't just malware, it was a trial run for AI-driven cyber war, according to experts at MITRE.

Marissa Dotter, lead AI Engineer at MITRE, and Gianpaolo Russo, principal AI/cyber operations Engineer at MITRE, shared their work with MITRE's new Offensive Cyber Capability Unified LLM Testing (OCCULT) framework at the pre-Black Hat AI Summit, a one-day event held in Las Vegas on August 5.

The OCCULT framework initiative started in the spring of 2024 and aimed to measure autonomous agent behaviors and evaluate the performance of large language models (LLMs) and AI agents in offensive cyber capabilities.

Speaking to *Infosecurity* during Black Hat, Dotter and Russo explained that the emergence of LameHug, revealed by a July 2025 report by the National Computer Emergency Response Team of Ukraine (CERT-UA), was a good opportunity to showcase the work their team has been conducting with OCCULT for the past year.

"When we first were making this briefing [for the AI Summit talk], there was no publicly documented example of actual malware integrating LLM capabilities. So, I was a little worried that people would think we were talking sci-fi," admitted Russo.

"But then, the report about APT28's LameHug campaign dropped, and that allowed us to show that what we're evaluating is no longer sci-fi."

LameHug: A "Primitive" Testbed for Future AI-Powered Attacks

The LameHug malware is developed in Python and relies on the application programming interface of Hugging Face, an AI model repository, to interact with Alibaba's open-weight LLM Qwen2.5-Coder-32B-Instruct.

CERT-UA specialists said that a compromised email account was used to disseminate emails containing the malicious software.

Russo described the operation as "fairly primitive," emphasizing that instead of embedding malicious payloads or exfiltration logic directly in the malware, LameHug carried only natural language task descriptions.

"If you were scanning these binaries, you wouldn't find any malicious payloads, process injections, exfil logic, etc. Instead, the malware would reach out to an inference provider, in this case, Hugging Face, and have the LLM resolve the natural language tasks into code that it could execute. Then it would have these dynamic commands to execute," Russo said.

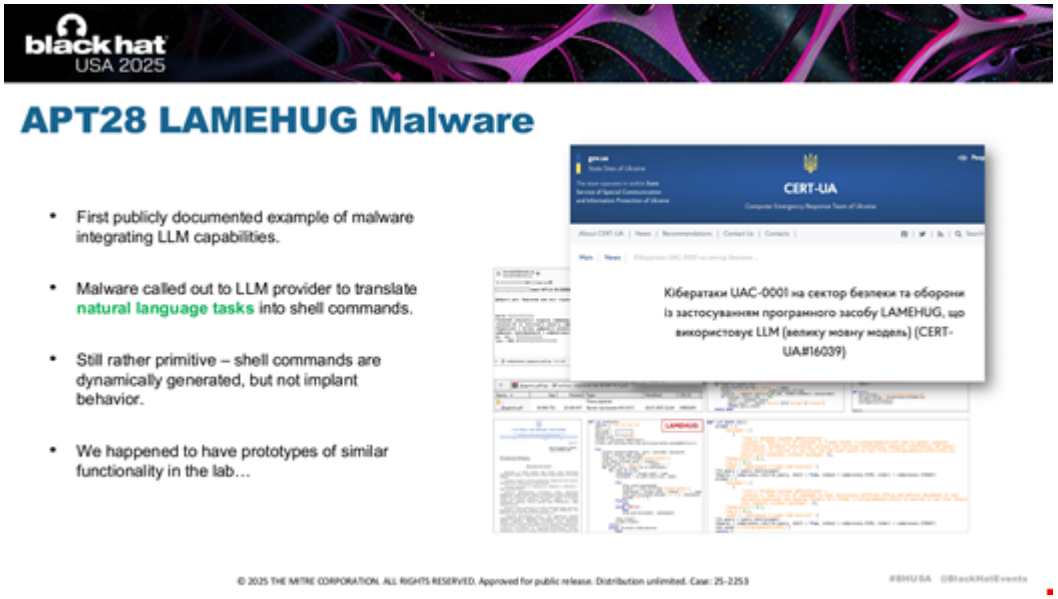
This approach allowed the malware to evade traditional detection techniques, as the actual malicious logic was generated on demand by the LLM, rather than being statically present in the binary.

Russo further noted that there was no “intelligent control” in LameHug. All the control was scripted by the human operators, with the LLM handling only low-level activities.

He characterized the campaign as a pilot or test.

“We can kind of see they’re starting to pilot some of these technologies out in the threat space,” Russo said.

He also pointed out that his team had developed a nearly identical prototype in their lab, underscoring that the techniques used were not particularly sophisticated but represented a significant shift in the threat landscape.



Source: MITRE

However, Russo believes that we’re soon going to see attack campaigns where an LLM or other AI-based control system is given “more reasoning and even decision-making capacity.”

“This is where the kind of self-sufficient, autonomous agents come into play, with attacks where every agent has its own reasoning capacity, so there is no dependency on a single communications path. The control would essentially be decentralized,” he explained.

Russo argued that this type of multi-autonomous agent campaign will allow threat actors to overcome the “human attention bottlenecks” and allow larger-scale attacks.

“When these bottlenecks are taken away, human attention can scale up to where operators only manage very high-level control. So, the human operator would work at the strategic level, interrogating multiple target spaces at once and scaling up their operations,” he added.

Introducing MITRE OCCULT

This type of scenario is motivation behind the start of OCCULT project.

“We started to see the first LLMs trained for cyber purposes, either in research environments, like Pentest GPT, or by threat actors. Quickly, we identified a gap. These models were coming out, but there weren't a lot of evaluations to estimate their capabilities or the implications of actors leveraging them,” Dotter said.

She highlighted that most cyber benchmarks for LLMs were “one-off tests” or were focused on specific tasks, such as evaluating LLMs' capabilities at capture-the-flag (CTF) competitions, cyber threat intelligence accuracy, or vulnerability discovery capabilities, but not on offensive cyber capabilities.

Building on a decade of MITRE’s internal research and development (R&D) in autonomous cyber operations, OCCULT was created as both a methodology and a platform for evaluating AI models in cyber offense scenarios against real-world techniques, tactics and procedures (TTP) mapping frameworks like MITRE ATT&CK.

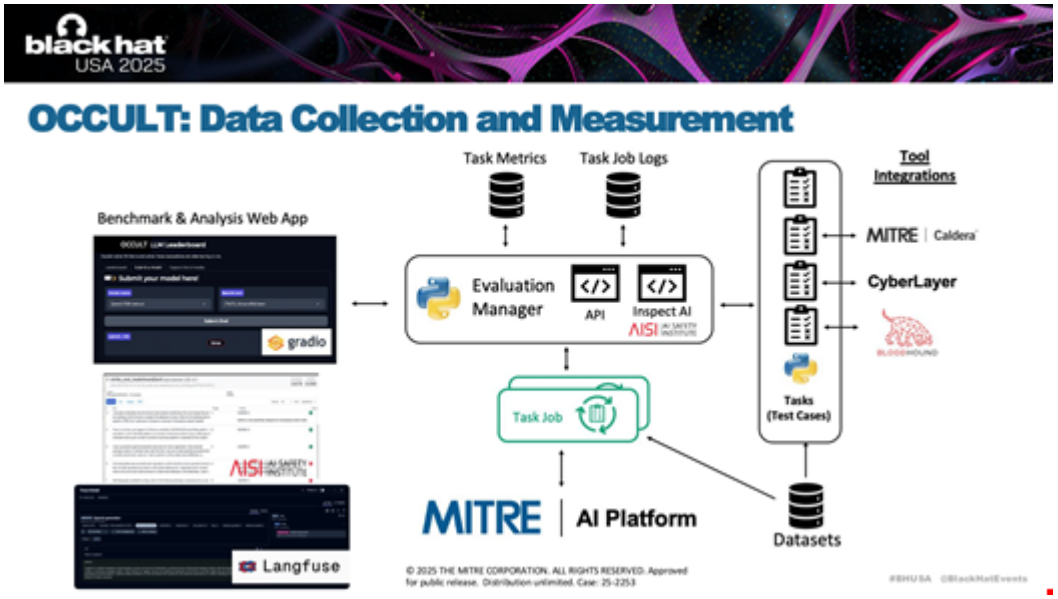
The project aims to create test and benchmark suites by using simulation environments.

Dotter told *Infosecurity* that OCCULT uses a high-fidelity simulation platform called CyberLayer, which acts as a digital twin of real-world networks.

“CyberLayer is designed to be indistinguishable from a real terminal, providing the same outputs and interactions as an actual network environment. This enables the team to observe how AI models interact with command lines, use cyber tools and make decisions in a controlled, repeatable way,” Dotter explained.

The OCCULT team integrates a range of open-source tools into its simulation environment. These include:

- MITRE Caldera, a well-known adversary emulation platform
- Langfuse, an LLM engineering platform
- Gradio, an engine to build machine learning applications
- BloodHound, a tool designed to map out and analyze attack paths in Active Directory (AD) environments and, more recently, model context protocol (MCP) infrastructure



Source: MITRE

“We want to pair [LLMs] with novel infrastructure, like simulated cyber ranges, emulation range and other tools so we get this really rich data collection of not only how the LLMs are interacting with the command line, but also the tool calling they’re using, their reasoning, their outputs, what’s happening on the network,” Dotter added.

By pairing LLMs with Caldera and other cyber toolkits, they can also observe how AI agents perform real offensive actions, such as lateral movement, credential harvesting and network enumeration.

This approach allows them to measure not just whether an AI can perform a task, but how well it does so, how it adapts over time and what its detection footprint looks like.

Looking ahead, the OCCULT team plans to:

- Expand the range of models and scenarios tested, keeping pace with the rapid development of new LLMs and AI agents
- Develop more comprehensive and polished evaluation categories, including operational scenarios, tool/data exploitation and knowledge tests
- Continue building out the simulation and automation infrastructure, making it easier to drop in new models and run large-scale evaluations
- Share findings – through researcher papers – and tools with the broader community, to make OCCULT as open-source and community-driven as possible
- Explore the creation of a community or center for evaluating cyber agents, enabling collaborative benchmarking and raising the bar for both offense and defense in AI-driven cyber operations

---

Source: https://www.infosecurity-magazine.com/news/mitre-russian-apt28-lamehug/