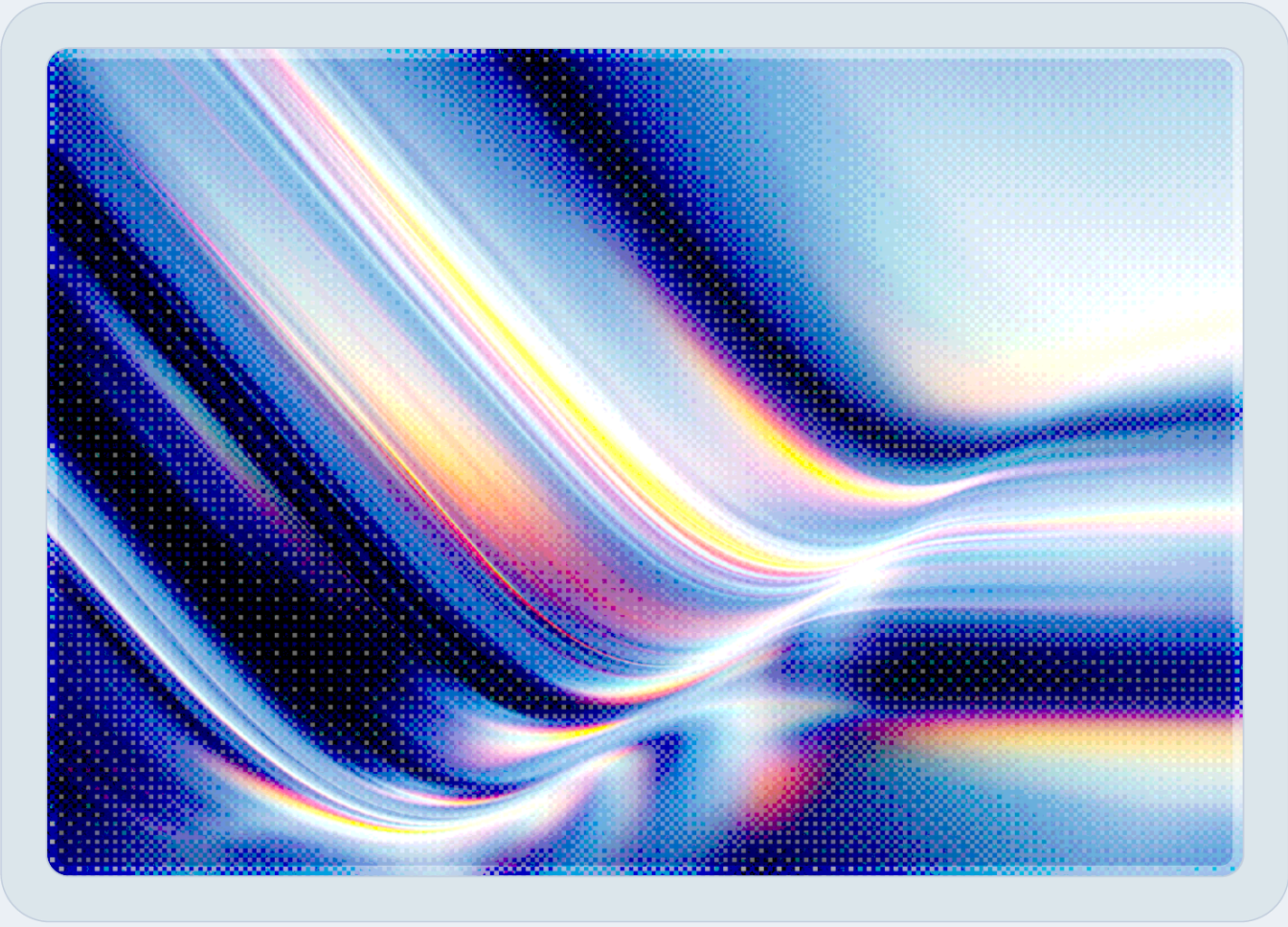
CSAI Foundation | Cloud Security Alliance

LLMjacking Evolved: Stolen AI Compute as Offensive Infrastructure

From Black Market Resale to Autonomous Attack Pipelines

2026-06-20

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- The June 2026 Sysdig finding signals that LLMjacking may be crossing a strategic threshold: for the first time, stolen AI compute has been documented not merely as resold infrastructure but as the reasoning engine wired into an autonomous offensive pipeline capable of scanning, exploiting, and compromising targets without human operators in the loop.
- On June 12, 2026, Sysdig's Threat Research Team captured the first confirmed instance of an attacker using a misconfigured Ollama model server as the reasoning engine for a multi-stage, VAPT-style attack pipeline—the first documented use of compromised AI inference as the cognitive core of an autonomous offensive operation.
- Approximately 175,000 Ollama instances are publicly accessible on the internet with no authentication by default, and CVE-2026-7482 ("Bleeding Llama") allows unauthenticated attackers to extract API keys, environment variables, and conversation data from any unpatched server using three API calls [1].
- The March 2026 LiteLLM supply chain compromise (CVE-2026-33634, CVSS 9.4) demonstrated that a single malicious dependency in an AI gateway library can simultaneously expose an organization's entire portfolio of AI provider credentials [3].
- Organizations must treat AI model-serving infrastructure—Ollama, LM Studio, LangServe, OpenWebUI, and AI gateway proxies—as high-value targets requiring the same zero-trust hardening applied to production databases and identity providers.

Background

The term "LLMjacking" was coined by Sysdig's Threat Research Team in May 2024 to describe a then-novel attack class: compromising cloud credentials or API keys specifically to consume AI inference capacity at the victim's expense [4]. The earliest documented campaigns were financially motivated and operationally simple. Attackers scraped API keys from public code repositories and misconfigured cloud storage, validated them against a roster of providers—OpenAI, Anthropic, Azure OpenAI, AWS Bedrock, Google Cloud Vertex AI, Mistral, and others—and resold the validated access through reverse-proxy networks on underground forums. Sysdig's original analysis established that a single Claude 2.x victim account could generate $46,080 in inference costs per day; attacks targeting Claude 3 Opus exceeded

$100,000 per day [4][17]. The financial asymmetry between the attacker's acquisition cost and the victim's bill was stark: validated access credentials sold through Discord and Telegram communities at prices that made even high-volume victims vastly profitable to exploit [4].

Through 2025 and into early 2026, the criminal ecosystem matured rapidly. Operation Bizarre Bazaar, documented jointly by Sysdig and Pillar Security Research in February 2026, provided the first fully attributed, end-to-end picture of LLMjacking as an organized industry [6][7]. The campaign, active between December 2025 and January 2026, recorded more than 35,000 attack sessions against exposed AI infrastructure. Attackers exploited remote code execution vulnerabilities in outdated web application frameworks to obtain cloud credentials, then validated and triaged those credentials against more than 30 LLM providers before listing access for resale through a commercial marketplace called silver[.]inc [6][7]. The campaign represented a departure from ad hoc opportunism toward systematic supply chain criminalization: reconnaissance, credential validation, quality-based victim triage, and commercial-grade marketplace distribution operated as discrete, coordinated stages. Sysdig's broader 2026 threat intelligence reporting documented a 376% increase in credential theft specifically targeting AI services between the fourth quarter of 2025 and the first quarter of 2026 [6][8].

The June 2026 development that forms the central subject of this analysis suggests LLMjacking may be evolving beyond its origins as primarily an infrastructure-abuse problem in which stolen compute is sold to third parties. The captured incident signals an emerging pattern in which stolen compute is used directly by adversaries to reason through attack sequences, generate exploits, and autonomously pursue targets—a trajectory that, if it continues, would transform compromised AI infrastructure from a commodity into a weapon.

Security Analysis

The Agentic Turn: From Resale to Weaponization

On June 12, 2026, Sysdig's Threat Research Team captured an attacker using a publicly exposed, unauthenticated Ollama model server not for resale or personal use but as the cognitive core of a multi-stage offensive pipeline [1]. Unlike the black-market monetization documented in Operation Bizarre Bazaar, this actor integrated unauthenticated model inference into a software architecture designed to automate the complete attack lifecycle. The captured pipeline performed automated reconnaissance against target networks, mapped discovered services to known vulnerability signatures, synthesized proof-of-concept exploit code, and attempted command execution—with the language model generating structured routing decisions at each stage, directing subsequent actions based on parsed outputs from the previous step. The design followed the same structure as legitimate vulnerability

assessment and penetration testing (VAPT) tooling, suggesting the actors were either former security professionals or had adapted open-source offensive frameworks to use a stolen AI backend in place of a licensed one [1].

The targets captured during the observation window were non-routable private addresses and known penetration testing lab environments, indicating the tool was still in refinement rather than active production deployment against real-world victims. That qualification should not reduce urgency: in prior threat generations—ransomware deployment timelines, initial access broker ecosystem development, and exploit kit distribution—the transition from prototype to operationalized capability has typically occurred faster than defenders anticipated, and there is no structural reason to expect this evolution will be slower. The infrastructure required for this capability—unauthenticated Ollama servers—is already abundant on the public internet.

This represents a qualitative shift in attacker economics. Traditional autonomous attack tooling requires the attacker to invest in purpose-built exploit frameworks, maintain hard-coded logic for each target type, and update that logic as targets evolve. An LLM-powered pipeline is inherently adaptive: it can process novel service banners, interpret unfamiliar error messages, and synthesize exploit strategies for previously unseen configurations in real time, all at the cost of inference tokens that the attacker is not paying for.

Exposed AI Serving Infrastructure as the Attack Surface

The June 2026 Sysdig incident was enabled by a structural property of the self-hosted AI serving ecosystem: Ollama, the most widely deployed open-source local model server, listens on port 11434 with no authentication enabled by default [2]. Researchers have catalogued approximately 175,000 publicly accessible Ollama instances spanning more than 130 countries [1]. Cisco Security documented the scope of Ollama internet exposure through Shodan analysis, observing that many operators deploy Ollama on cloud virtual machines without firewall rules blocking inbound access to the model serving port, effectively offering free AI inference to any scanner that finds the host [9].

CVE-2026-7482, disclosed in May 2026 with CVSS v3.1 scores ranging from 8.8 to 9.3 depending on the assessing organization (runZero assessed it at 9.1 [2]; SentinelOne at 8.8 [10]), compounded the exposure significantly. Nicknamed "Bleeding Llama," the vulnerability is a heap out-of-bounds read in Ollama versions prior to 0.17.1 affecting the `/api/create` endpoint, which accepts GGUF model files without enforcing bounds on tensor dimension metadata. An attacker submits a crafted GGUF file with tensor offset and size values that exceed the file's actual length, coercing the server into reading beyond the mapped memory region. The resulting memory contents can be exfiltrated by pushing the crafted model artifact to an attacker-controlled registry via `/api/push`. Neither endpoint requires authentication in the default upstream configuration. Leaked memory may contain API keys for

downstream provider integrations, environment variables, system prompts, and conversation data from concurrent sessions [2][10]. Separate scanning research conducted around the time of CVE-2026-7482's disclosure estimated the internet-facing Ollama population at several hundred thousand instances, reflecting different scanning methodologies and measurement windows from those used in the Sysdig June 2026 assessment.

The exposure problem extends beyond Ollama. A Kaspersky honeypot study published in May 2026 impersonated multiple common local AI serving stacks—Ollama, LM Studio, AutoGPT, LangServe, and text-generation-webui—and advertised a locally hosted large language model. Over one month, the honeypot recorded more than 113,000 probe requests from thousands of unique IP addresses, with 23% of traffic specifically probing for AI model serving capabilities rather than general web services [11]. OpenWebUI servers have also been observed hijacked for cryptomining, demonstrating that financially motivated actors are actively scanning for any AI infrastructure accessible without authentication [12]. The combined picture is of a largely unguarded attack surface: tens of thousands of AI serving endpoints deployed by developers, researchers, and small enterprises without the security disciplines—network segmentation, authentication enforcement, access logging—routinely applied to production databases and comparable infrastructure.

Supply Chain Compromise: Poisoning the AI Gateway

A second, structurally distinct threat vector emerged in March 2026 with the compromise of LiteLLM, an LLM gateway library with more than 95 million monthly PyPI downloads [3][13]. On March 24, 2026, malicious versions 1.82.7 and 1.82.8 of the `litellm` package were uploaded to the Python Package Index. The attacker, attributed to a criminal group tracked as TeamPCP, used a `.pth` file—a legitimate Python interpreter mechanism that auto-executes code at interpreter startup without requiring an explicit import—to deploy a three-stage payload: credential harvesting, Kubernetes lateral movement, and a persistent remote code execution backdoor [13][14]. The compromised versions were live for approximately 40 minutes, but a window of that duration is typically sufficient for automated dependency pipelines to ingest a new package version.

The LiteLLM attack is architecturally significant because LiteLLM functions as a credential aggregation layer: a single deployment typically holds API keys for every AI provider an organization uses—OpenAI, Anthropic, Google, Cohere, Mistral, and others—along with any API keys passed through for routing. Compromising LiteLLM therefore does not yield one provider credential but potentially the entire organizational AI credential portfolio in a single operation. The attack was assigned CVE-2026-33634 with a CVSS severity of 9.4 [3]. TeamPCP conducted a coordinated campaign that simultaneously

trojanized the Trivy container security scanner and the Checkmarx static analysis tool through analogous PyPI supply chain techniques, suggesting a deliberate strategy of targeting security and AI tooling to maximize the blast radius of a single supply chain operation [14].

The LiteLLM incident underscores a pattern security teams must internalize: the AI infrastructure stack—model servers, gateway proxies, orchestration frameworks, and the Python packages that compose them—has become as high-value a target as identity providers and secrets management systems, while typically receiving significantly less security scrutiny.

The Composite Threat Picture

These three vectors—hijacked open model servers, CVE-exploited AI serving infrastructure, and supply chain compromise of AI gateways—are not independent phenomena. They interact to create a composite threat environment. An attacker who compromises a LiteLLM deployment obtains API keys that can be validated, quality-triered, and resold through the Operation Bizarre Bazaar model. An attacker who discovers a CVE-2026-7482-vulnerable Ollama server extracts API keys that provide access to cloud-hosted frontier models at victim expense. An attacker building the June 2026-style autonomous offensive pipeline needs free AI inference as a prerequisite; the 175,000 unauthenticated Ollama instances identified by Sysdig provide exactly that.

The CrowdStrike 2026 Global Threat Report reported that adversaries compromised AI tools at more than 90 organizations in the preceding year, with the average eCrime breakout time falling to 29 minutes—with the fastest observed breakout at 27 seconds—and AI-enabled adversary operations increasing 89% year over year [15]. Flashpoint observed more than 11.1 million machines infected with infostealers in 2025, generating an inventory of 3.3 billion compromised credentials and cloud tokens available to criminal buyers [16]. AI API keys represent a particularly high-value subset of that inventory: they offer both immediate financial return through compute arbitrage and, as the June 2026 Sysdig finding demonstrates, a potential force multiplier for subsequent offensive operations.

Recommendations

Immediate Actions

Organizations running Ollama, LM Studio, LangServe, OpenWebUI, or similar self-hosted AI serving infrastructure should audit network exposure immediately. Any model serving endpoint accessible from the public internet without authentication and without explicit organizational intent to offer public inference represents an unacceptable risk given current attacker scanning activity. CVE-2026-7482-

affected Ollama deployments (versions prior to 0.17.1) should be patched or taken offline without delay, as the vulnerability enables unauthenticated memory extraction requiring no credentials and only three API calls.

Organizations using LiteLLM should audit installed versions and deployment pipelines for CVE-2026-33634 exposure. Any system that downloaded litellm versions 1.82.7 or 1.82.8 between March 24, 2026, and the time of package removal should be treated as potentially compromised: rotate all AI provider API keys, review Kubernetes service account permissions, and scan for persistence mechanisms consistent with .pth-based payload delivery.

AI provider API keys stored in environment variables, configuration files, or secrets managers should be rotated on a defined schedule and scoped with least-privilege access. Provider-specific usage alerts—available natively in the AWS Bedrock, Azure OpenAI, and Anthropic consoles—should be configured to trigger on anomalous inference volume or unexpected model types.

Short-Term Mitigations

AI model serving infrastructure should be placed behind network controls equivalent to those applied to internal databases: no public internet exposure, authentication enforced at the API layer using short-lived tokens issued via OIDC or OAuth2, and network segmentation preventing model servers from reaching external credential stores directly. The Kaspersky honeypot study recommends treating private model-serving infrastructure with the same hardening applied to production servers [11], a posture that most current deployments do not meet.

Supply chain risk management processes should extend to AI-specific dependencies. LiteLLM, LangChain, LlamaIndex, and similar orchestration libraries should be subject to the same software composition analysis, version pinning, and dependency integrity verification as production application dependencies. The TeamPCP campaign's simultaneous trojanization of Trivy and Checkmarx alongside LiteLLM suggests that security tooling itself is a deliberate target, and security teams should not assume that their vulnerability scanning and static analysis tools are inherently trustworthy without integrity verification of the installed artifacts.

Strategic Considerations

The June 2026 Sysdig finding signals that AI infrastructure must be incorporated into threat modeling not merely as an asset to protect but as a potential capability that adversaries may use against organizations' other assets. Security architects designing AI deployments should consider the adversarial

use case: if an attacker gained unauthenticated access to the organization's model serving layer, what offensive capability would they acquire? Threat models should account for both the data exposed by the AI system and the reasoning capability it could provide to an autonomous attack pipeline.

CSA's prior analysis of LLMjacking's black market dimension, published through CSA Labs in March 2026, focused primarily on credential theft and financial impact [17]. The agentic evolution documented here requires security programs to extend their response frameworks to address AI as an active component of the adversary's operational toolkit. Detection strategies should include monitoring for inference patterns inconsistent with organizational AI use cases—high-volume automated prompt sequences, programmatic error recovery patterns, and structured output formats characteristic of tool-calling pipelines—in addition to traditional cost anomaly alerting.

CSA Resource Alignment

The threat described in this research note maps across multiple layers of the MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome) agentic AI threat modeling framework developed by CSA [18]. The June 2026 offensive pipeline exploits vulnerabilities at MAESTRO's Layer 1 (Foundation Models, where unauthenticated access to hosted models occurs), Layer 3 (Agent Frameworks, where the autonomous attack pipeline logic operates), and Layer 4 (Deployment and Infrastructure, where misconfigured Ollama servers and exposed gateway proxies reside). MAESTRO's emphasis on cross-layer attack paths is directly applicable: the LiteLLM supply chain compromise, for example, creates exposure at both the infrastructure layer (credential theft) and the agent framework layer (where the persistent RCE backdoor could enable future manipulation of routing and model selection logic).

The AI Controls Matrix (AICM) v1.0, CSA's AI security control framework, addresses several of the control gaps this note highlights. AICM domains covering AI supply chain security, identity and access management for AI systems, and network security for AI workloads are directly relevant to the credential theft and exposure vectors described. Organizations using AICM as an audit baseline should specifically examine controls governing third-party AI component integrity and model serving endpoint authentication. CSA's AI-CAIQ (AI Consensus Assessment Initiative Questionnaire) provides a structured mechanism for vendors to attest to their posture against these controls, and procurement teams should require LLM gateway and model serving vendors to complete AI-CAIQ assessments.

The CSA STAR for AI program offers an independent assurance pathway for organizations seeking third-party validation of their AI security posture against AICM controls. Given the supply chain dimension of the threats described here—where the risk originates not in the organization's own AI models but in the

libraries and proxies surrounding them—STAR for AI assessments should encompass the full AI infrastructure stack, including gateway libraries, orchestration frameworks, and self-hosted model serving endpoints.

CSA's Zero Trust guidance is directly applicable to the network exposure problem: the default-open posture of Ollama and similar tools is antithetical to zero trust principles, and organizations should enforce the principle that no model serving endpoint is trusted by network position alone. CSA's runtime detection guidance for AWS Bedrock and SageMaker workloads provides a practical reference for instrumentation and alerting patterns that can be adapted to self-hosted infrastructure [19].

References

[1] Sysdig Threat Research Team. "LLMjacking Evolved: Attackers Are Using Stolen AI Compute to Build Offensive Agentic Tools." Sysdig, June 2026.

[2] runZero Research. "Ollama Vulnerability CVE-2026-7482: Find Impacted Assets." runZero, May 2026.

[3] LiteLLM. "Security Update: Suspected Supply Chain Incident." LiteLLM Official Blog, March 2026.

[4] Sysdig Threat Research Team. "LLMjacking: Stolen Cloud Credentials Used in New AI Attack." Sysdig, May 2024.

[5] Infosecurity Magazine. "New 'LLMjacking' Attack Exploits Stolen Cloud Credentials." Infosecurity Magazine, 2024.

[6] Pillar Security Research. "Operation Bizarre Bazaar: First Attributed LLMjacking Campaign with Commercial Marketplace Monetization." Pillar Security, February 2026.

[7] SecurityWeek. "LLMs Hijacked, Monetized in 'Operation Bizarre Bazaar'." SecurityWeek, February 2026.

[8] Sysdig Threat Research Team. "LLMjacking: From Emerging Threat to Black Market Reality." Sysdig, 2026.

[9] Cisco Security. "Detecting Exposed LLM Servers: A Shodan Case Study on Ollama." Cisco Blogs, 2025.

[10] SentinelOne. "CVE-2026-7482: Ollama Information Disclosure Vulnerability." SentinelOne Vulnerability Database, May 2026.

[11] Kaspersky. "LLMjacking: What These Attacks Are, and How to Protect AI Servers." Kaspersky Blog, May 2026.

[12] Cybernews. "Hackers Turned OpenWebUI AI Servers Into Crypto Mining Zombies." Cybernews, 2026.

[13] Kaspersky. "Trojanization of Trivy, Checkmarx, and LiteLLM Solutions." Kaspersky Blog, 2026.

[14] Securelist (Kaspersky). "Why Is the LiteLLM AI Gateway Compromise So Dangerous?" Securelist, 2026.

[15] CrowdStrike. "2026 CrowdStrike Global Threat Report." CrowdStrike, February 2026.

[16] HSToday. "2026 Global Threat Intelligence Report Highlights Rise in Agentic AI Cybercrime." HSToday, 2026.

[17] Cloud Security Alliance AI Safety Initiative. "LLMjacking: AI Model Hijacking Reaches Black Market Scale." CSA Labs, March 2026.

[18] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 2025.

[19] Cloud Security Alliance. "Securing AI Workloads in AWS: Why Bedrock and SageMaker Need Runtime Detection and AI-Powered Response." CSA Blog, June 2026.