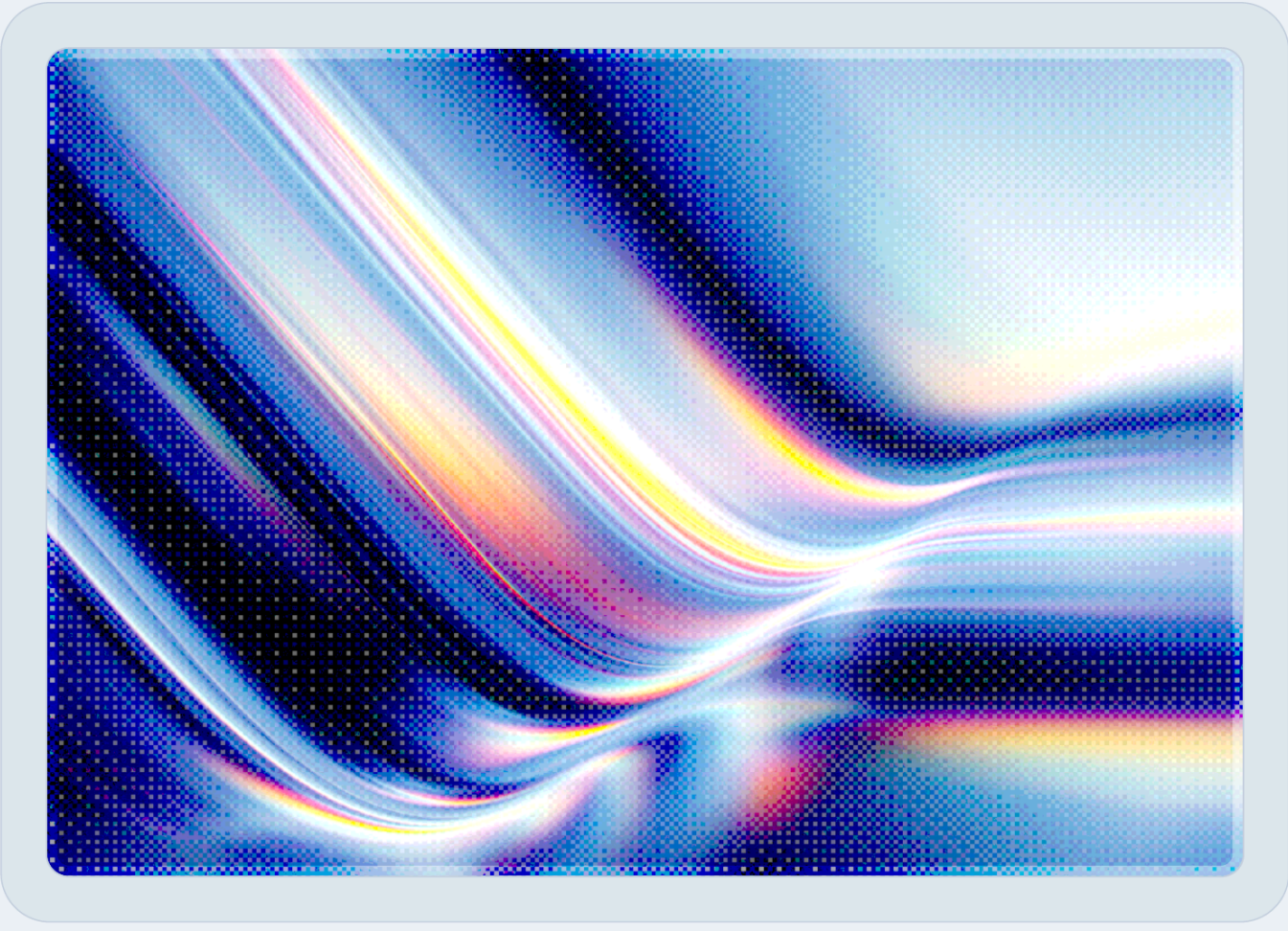
CSAI Foundation | Cloud Security Alliance

LLMjacking Evolved: Stolen AI Compute as Offensive Infrastructure

From Credential Theft and Resale to Autonomous Offensive Agentic Pipelines

2026-07-01

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

- LLMjacking — the theft of cloud AI credentials to consume inference capacity at the victim's expense — has evolved in under two years from opportunistic API key scraping to industrial-scale criminal infrastructure, with Sysdig documenting a 376% increase in credential theft targeting AI services between Q4 2025 and Q1 2026 [1].
- Operation Bizarre Bazaar, a joint Sysdig and Pillar Security Research attribution campaign spanning December 2025 to January 2026, documented the first fully commercial LLMjacking marketplace: a three-stage supply chain of automated scanning, credential validation, and resale through a dedicated storefront, generating 35,000 attack sessions across 30+ LLM providers in approximately 36 days [2].
- Researchers from SentinelOne SentinelLABS and Censys identified 175,000 publicly exposed Ollama AI server instances across 130 countries in early 2026, with nearly half configured to support tool-calling capabilities including code execution, external API access, and file system interaction — providing attackers with ready-made infrastructure to exploit [3].
- In June 2026, Sysdig's Threat Research Team documented a qualitative shift: a threat actor was no longer reselling stolen AI compute but wiring it directly as the reasoning engine into an automated offensive pipeline (designated VAPT) capable of service fingerprinting, vulnerability matching, proof-of-concept generation, SQL injection crafting, and privilege escalation [4].
- By late January 2026, 60% of observed attack traffic against exposed AI endpoints had shifted from pure compute theft toward Model Context Protocol (MCP) reconnaissance — probing connected file systems, databases, shell integrations, and Kubernetes clusters — suggesting attackers are mapping attached capabilities as much as consuming inference [2].
- Defenses must address the entire attack chain: hardening AI infrastructure exposure, enforcing API key governance with short-lived scoped tokens, monitoring for anomalous inference volume, and treating unauthenticated AI endpoints with the same urgency as exposed cloud storage buckets.

Background

The term "LLMjacking" was coined by the Sysdig 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 [5]. The initial documented campaign exploited CVE-2021-3129, a CVSS 9.8 remote code execution vulnerability in Laravel's Ignition debug mode, to compromise a web server, pivot to its embedded cloud credentials, and invoke AI services across ten different cloud providers in a single operation. Sysdig estimated that running a Claude 2.x-class model at full capacity could impose costs of roughly $46,000 per day on the compromised account — a figure that has since risen substantially as attackers have shifted to higher-capability models [5, 6].

The underlying economics of LLMjacking are straightforward and have proven durable. Legitimate enterprise AI inference is expensive; premium frontier models charge tens of dollars per million tokens. A stolen credential resells at 40 to 60 percent below legitimate pricing [2], sustaining the economics of a credential-resale industry. The victim, meanwhile, discovers the attack only when an unexpectedly large cloud bill appears — or when their API key is revoked for terms-of-service violations before they realize the account was compromised.

Attackers diversified their credential acquisition methods rapidly. Code repositories remain a primary source: developers who accidentally commit API keys or embed them in environment files accessible to public repositories continue to provide a large, automatically refreshed supply of working credentials. Phishing campaigns targeting AI platform accounts, infostealers harvesting browser-stored API tokens, and misconfigured cloud storage exposing `.env` files or `~/.config` credential files round out the acquisition pipeline. Flashpoint's 2026 Global Threat Intelligence Report documented more than 11.1 million machines infected with infostealers during 2025, generating an inventory of 3.3 billion compromised credentials and cloud tokens circulating in criminal markets — a pool from which AI API keys represent an increasingly valued subset [7].

The threat grew in both scale and organization between mid-2024 and the end of 2025. Attackers moved from manually testing individual stolen keys to operating automated validation pipelines that discarded invalid keys and categorized valid ones by provider, model tier, and remaining quota before listing them for resale. This automation compressed the time between credential theft and active exploitation, reducing the window in which victims could rotate keys before incurring costs [2].

Security Analysis

Operation Bizarre Bazaar: LLMjacking as an Organized Industry

Operation Bizarre Bazaar, jointly documented by Sysdig and Pillar Security Research in a February 2026 report, provided the first fully attributed end-to-end picture of LLMjacking as an organized criminal enterprise [2]. The campaign, active between December 2025 and January 2026, was traced to a threat actor designated Hecker (also known as Sakuya and LiveGamer101), who operated a commercial marketplace called silver[.]inc that listed access to validated AI API credentials across more than 30 providers.

The operation's structure exhibited the hallmarks of professionalized criminal infrastructure. A distributed scanning layer continuously probed the internet for exposed AI endpoints — Ollama instances on port 11434, OpenAI-compatible APIs on port 8000, unauthenticated MCP servers, and development environments with public IP addresses. Discovered credentials and endpoints passed through an automated validation pipeline that confirmed access, measured available quota, and assessed model tier before forwarding them to the marketplace layer. Silver[.]inc offered buyers point-and-click access to validated AI inference at below-market rates, with the stolen account absorbing all costs. Researchers' honeypots captured 35,000 attack sessions during the campaign's operational window, averaging 972 attacks per day [2].

The campaign's MCP reconnaissance behavior deserves particular attention. By late January 2026, 60% of attack traffic observed by Pillar Security researchers had shifted from consume-and-resell patterns toward systematic probing of what tools and integrations were attached to exposed AI endpoints [2]. MCP servers connected to model instances were interrogated for their capability manifests: which file system paths they could access, what database connections they exposed, whether they had shell execution or Kubernetes API access, and what cloud service integrations were available. This reconnaissance activity may not generate inference costs and may not surface in API usage anomaly alerts — distinguishing it from active model invocation — though detection effectiveness depends on the monitoring architecture of the specific deployment. An attacker probing MCP capability manifests is assessing the attack surface adjacent to a compromised model server, mapping what tools and integrations are available before invoking the model itself.

The Exposed Ollama Attack Surface

SentinelOne SentinelLABS and Censys researchers identified 175,000 publicly exposed Ollama AI server instances across 130 countries in research published in early 2026 [3]. Ollama is an open-source framework for running large language models locally or on self-managed infrastructure; it is widely used in enterprise development environments, research labs, and cloud deployments where operators want to run models without routing requests through external APIs. The server defaults to unauthenticated access on port 11434 — a reasonable default for local development but a significant exposure when instances are reachable from the public internet, typically requiring only a single configuration change to bind Ollama to a public interface rather than localhost.

The geographic distribution of exposed instances spanned cloud and residential networks, with China, the United States, Germany, France, and South Korea representing the largest concentrations. Critically, nearly half of the observed hosts were configured with tool-calling capabilities, meaning the exposed model instance was not merely available for chat completion but could execute code, read and write files, call external APIs, and interact with connected systems through whichever tools had been wired into the deployment [3]. An attacker who discovers and connects to one of these instances obtains not just free inference capacity but a model with hands — a ready-made agent with access to whatever the host machine has connected to it.

The scale of this exposure creates a persistent reservoir for LLMjacking operations that requires no credential theft at all. Exposed Ollama instances require no authentication, no stolen key, and no account takeover. Operators conducting the reconnaissance phase of an offensive campaign can simply query exposed endpoints for their model lists and capability manifests, then select those best suited to their operational requirements.

From Resale to Reasoning Engine: The VAPT Discovery

On June 12, 2026, Sysdig's Threat Research Team documented the first confirmed instance of stolen AI compute being used as the reasoning engine in an automated multi-stage offensive pipeline [4]. A threat actor had located a misconfigured Ollama instance and, rather than reselling access to it or using it for AI-assisted content generation, had integrated it as the decision-making layer in a framework the researchers designated VAPT.

The VAPT framework observed in the June 2026 incident was not a research proof of concept. Sysdig attributed four separate operational sessions to the same actor based on the distinctive pipeline structure, shared private benchmark targets, and consistent network origin. The framework's recorded workflows encompassed service fingerprinting and port scanning, automated vulnerability matching against known CVE inventories, web application reconnaissance, proof-of-concept exploit generation,

SQL injection payload crafting, secret extraction from exposed configuration files, and privilege escalation sequences. The model acted as the decision-making layer at each stage: interpreting scan results, selecting applicable vulnerability patterns, generating exploit code, and determining whether to continue or pivot based on execution outcomes.

Two days later, on June 14, the actor directed VAPT against the 10.129.0.0/16 address range — the subnet used by HackTheBox penetration-testing laboratory environments [4]. The choice of a benchmark range is significant. Legitimate security researchers use HackTheBox machines to develop and validate offensive techniques against controlled targets before applying them elsewhere. The actor's use of the same benchmark range, combined with in-place rewrites to the pipeline's code between sessions, indicates active development and capability evaluation rather than operational deployment at scale. The operator was using stolen AI compute as the free iteration engine for building and improving an autonomous hacking tool, with each run of the VAPT pipeline generating results that could inform the next revision.

This pattern has the potential to substantially alter the economics of offensive AI tooling. Developing a capable autonomous exploitation pipeline against a frontier model API would, at retail prices, cost the developer directly in proportion to inference consumed during testing and refinement. By routing that development work through a stolen or unauthenticated Ollama instance, the developer externalizes that cost entirely. The victim infrastructure pays for the attacker's research and development.

The Escalating Risk of Autonomous Exploitation

The June 2026 VAPT observation is the first field-documented instance of stolen AI compute being used as an autonomous offensive reasoning engine, but the capability it represents has been anticipated in research for some time. A 2024 academic study demonstrated that a capable language model given a vulnerability description could autonomously exploit 87% of a set of one-day vulnerabilities without additional human guidance, identifying and chaining exploit components faster than human analysts working the same targets [8]. The VAPT pipeline suggests that threat actors are now constructing real-world systems that actualize this capability, not as academic exercises but as operational tools under active development.

The shift from human-operated exploitation to model-operated exploitation has concrete implications for defenders. Human attackers are rate-limited by cognitive bandwidth: they can work on a small number of targets simultaneously and move through the kill chain at human speed. An agentic offensive pipeline backed by a language model is constrained primarily by network bandwidth and the inference capacity available to it. A threat actor who brings such a pipeline to operational maturity could, in

principle, run exploitation campaigns against many targets in parallel, with the model making tactical decisions at each node without requiring human input between steps. The VAPT observation suggests this trajectory is underway, even if current deployments remain in active development.

Microsoft's January 2025 civil lawsuit against a group of threat actors illustrated one direction this commercialization can take [9]. When Microsoft publicly attributed the defendants as Storm-2139 the following month, the complaint described a "hacking-as-a-service" scheme built on stolen Azure OpenAI Service API keys: a custom reverse proxy routing customer requests through Cloudflare tunnels to obscure the underlying stolen credentials, with AI safety filters bypassed, giving buyers capabilities that the platform's terms of service prohibit [13]. Microsoft cited violations of the Computer Fraud and Abuse Act, the Digital Millennium Copyright Act, and RICO statutes, reflecting the legal weight the company assigned to AI credential theft and the downstream enablement of safety-bypassed content generation.

Recommendations

Immediate Actions

Organizations running any self-hosted AI inference infrastructure should audit network exposure as the first priority. Ollama instances, locally-deployed model servers, and any OpenAI-compatible API should be verified to require authentication and should be inaccessible from public internet addresses. This applies to development environments and staging deployments as well as production; Operation Bizarre Bazaar demonstrated that attackers systematically target all three. Firewall rules restricting AI server ports (11434 for Ollama, 8000 for generic OpenAI-compatible APIs) to authorized source addresses should be deployed immediately where they do not already exist.

API key hygiene requires immediate attention for all organizations using commercial AI APIs. Every API key in use should be inventoried, scoped to the minimum required model access, and assigned a rotation schedule. Keys that appear in code repositories — even private ones, even briefly — should be treated as compromised and revoked. Most major AI platforms support scoped keys that restrict access to specific models, impose per-key rate limits, and generate per-key billing visibility; these controls should be activated wherever available, as they transform an all-access stolen credential into a limited, monitorable one.

Short-Term Mitigations

Anomalous inference volume monitoring should be implemented for all AI API accounts and self-hosted inference endpoints. Baseline usage patterns vary significantly by organization and workload, but sudden spikes in token consumption — particularly outside business hours, against high-cost model tiers, or from IP addresses outside expected ranges — are reliable indicators of unauthorized use. All major AI providers expose per-key usage metrics through their APIs and dashboards; organizations should integrate these into existing SIEM or observability stacks and configure alerting thresholds appropriate to their baseline. For self-hosted infrastructure, inference request logging with source IP, model invoked, and token counts provides the visibility necessary to detect both LLMjacking and the MCP reconnaissance pattern observed in Operation Bizarre Bazaar.

MCP server deployments warrant particular scrutiny given the reconnaissance shift documented in that campaign. Any production AI deployment that exposes MCP integrations to model endpoints should audit what capabilities those integrations provide. MCP servers with access to file systems, databases, shell execution, or cloud provider APIs represent a significant expansion of the attack surface accessible through a compromised model endpoint. Organizations should apply the principle of least privilege to MCP tool configurations, ensuring that models are connected only to the capabilities required for their specific operational purpose and that sensitive integrations are not enabled in deployments accessible from less-trusted network segments.

Strategic Considerations

The progression from credential theft to autonomous offensive pipeline represents an expansion of the threat model that enterprise AI governance programs must incorporate. Traditional controls designed around managing human-operated API usage — rate limits, budget caps, usage dashboards — remain necessary but are no longer sufficient. The VAPT pattern demonstrates that an attacker who gains access to AI inference capacity through any means can use it to conduct operations that extend far beyond AI platform terms-of-service violations. Security teams should evaluate their AI infrastructure exposure using the same adversarial framing they apply to any externally-accessible compute: who could reach this, what could they do with it, and would we know?

Credential management for AI services should be treated as a distinct discipline rather than an extension of general secrets management. The economic characteristics of AI inference theft — high per-day costs, automated exploitation, broad provider targeting — create incentive structures that differ from traditional cloud credential abuse. Short-lived, workload-scoped tokens generated through service identity mechanisms (such as OIDC-based federation rather than static API keys) substantially reduce

the value of stolen credentials by limiting both their scope and their useful life. Where AI providers support federated identity mechanisms, organizations should prioritize their adoption over static key management.

Detection engineering for LLMjacking requires attention to patterns that existing cloud security tooling may not surface. Inference consumption anomalies differ from conventional compute abuse signals; security teams should work with their AI platform account representatives to understand what telemetry is available, what alerting thresholds are appropriate for their usage profiles, and how quickly suspicious activity can be flagged. The 60% shift toward MCP reconnaissance observed in late January 2026 attack traffic suggests that monitoring should extend beyond inference volume to include model endpoint query patterns — particularly requests that appear to be enumerating attached capabilities rather than performing productive tasks.

CSA Resource Alignment

LLMjacking in its current evolved form maps directly to several threat categories and control domains defined in CSA's MAESTRO agentic AI threat modeling framework [10]. MAESTRO Layer 4 (Deployment and Infrastructure) addresses the exposure of AI server endpoints and the authentication controls required to protect them — the foundational gap that both Operation Bizarre Bazaar and the exposed Ollama landscape exploit. The near-half of exposed Ollama instances configured with tool-calling capabilities compound this deployment-layer risk: an attacker who reaches an unauthenticated endpoint gains not just inference access but model-mediated access to whatever systems those tools connect to. Layer 3 (Agent Frameworks) addresses the threat of autonomous decision-making chains, which is the capability at the core of the VAPT offensive pipeline: a model making tactical decisions about exploitation without human operators in the loop.

The AI Controls Matrix (AICM) v1.0 provides control guidance applicable to each phase of the LLMjacking kill chain. Authentication and access controls in the Infrastructure and Identity domains directly address API key management, endpoint authentication requirements, and the preference for short-lived scoped credentials over static keys. The Monitoring and Logging domains address the inference volume telemetry and anomaly detection that defenders need to identify active LLMjacking operations. Supply chain security controls in the AICM address the risk that infostealer-harvested credentials enter attacker inventories through third-party tooling embedded in developer workflows rather than through direct compromise of AI accounts.

CSA's published guidance on Zero Trust architecture is particularly relevant to the self-hosted AI infrastructure exposure problem [11]. Zero Trust principles require that no network location be implicitly trusted: a model server on a corporate network segment is not secure by virtue of its location, and access to its inference API should require explicit authentication regardless of whether the requesting client is internal or external. The pattern of 175,000 unauthenticated Ollama instances suggests that this principle is not yet being systematically applied to AI infrastructure deployment — a gap that is primarily one of deployment practice rather than product design, since Ollama's default localhost binding is appropriate for local use and becomes a risk only when operators expose it to broader networks. Extending Zero Trust principles to AI endpoints — treating them as protected resources that require identity verification for every connection — would close a significant fraction of the current attack surface.

The CSA LLM Threats Taxonomy classifies the unauthorized use of AI inference capacity as an instance of the Denial of Service and Loss of Governance threat categories, both of which are represented in the LLMjacking pattern: victims lose governance over their AI accounts and, in high-volume attacks, may have service degraded by rate limits triggered by unauthorized consumption [12]. The taxonomy's coverage of AI supply chain security also encompasses the infostealer pathway through which many AI credentials are harvested — a reminder that the LLMjacking attack surface extends into endpoint security and developer toolchain hygiene, not only AI platform administration.

References

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

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

[3] SentinelOne SentinelLABS and Censys. "Silent Brothers: Ollama Hosts Form Anonymous AI Network Beyond Platform Guardrails." SentinelOne, January 2026.

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

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

[6] Auth0 by Okta. "LLMjacking and the Hidden Cost of a Stolen API Key." Auth0, 2025.

[7] Flashpoint Intelligence. "2026 Global Threat Intelligence Report." Flashpoint, 2026.

[8] R. Fang, R. Bindu, A. Gupta, Q. Zhan, D. Kang. "LLM Agents Can Autonomously Exploit One-Day Vulnerabilities." arXiv, 2024.

[9] TechCrunch. "Microsoft Accuses Group of Developing Tool to Abuse Its AI Service in New Lawsuit." TechCrunch, January 2025.

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

[11] Cloud Security Alliance. "Zero Trust Guidance for Critical Infrastructure." CSA, 2024.

[12] Cloud Security Alliance. "Large Language Model (LLM) Threats Taxonomy." CSA, 2024.

[13] Microsoft. "Disrupting a Global Cybercrime Network Abusing Generative AI." Microsoft On the Issues, February 2025.

© 2026 Cloud Security Alliance. AI-assisted Rapid Research.