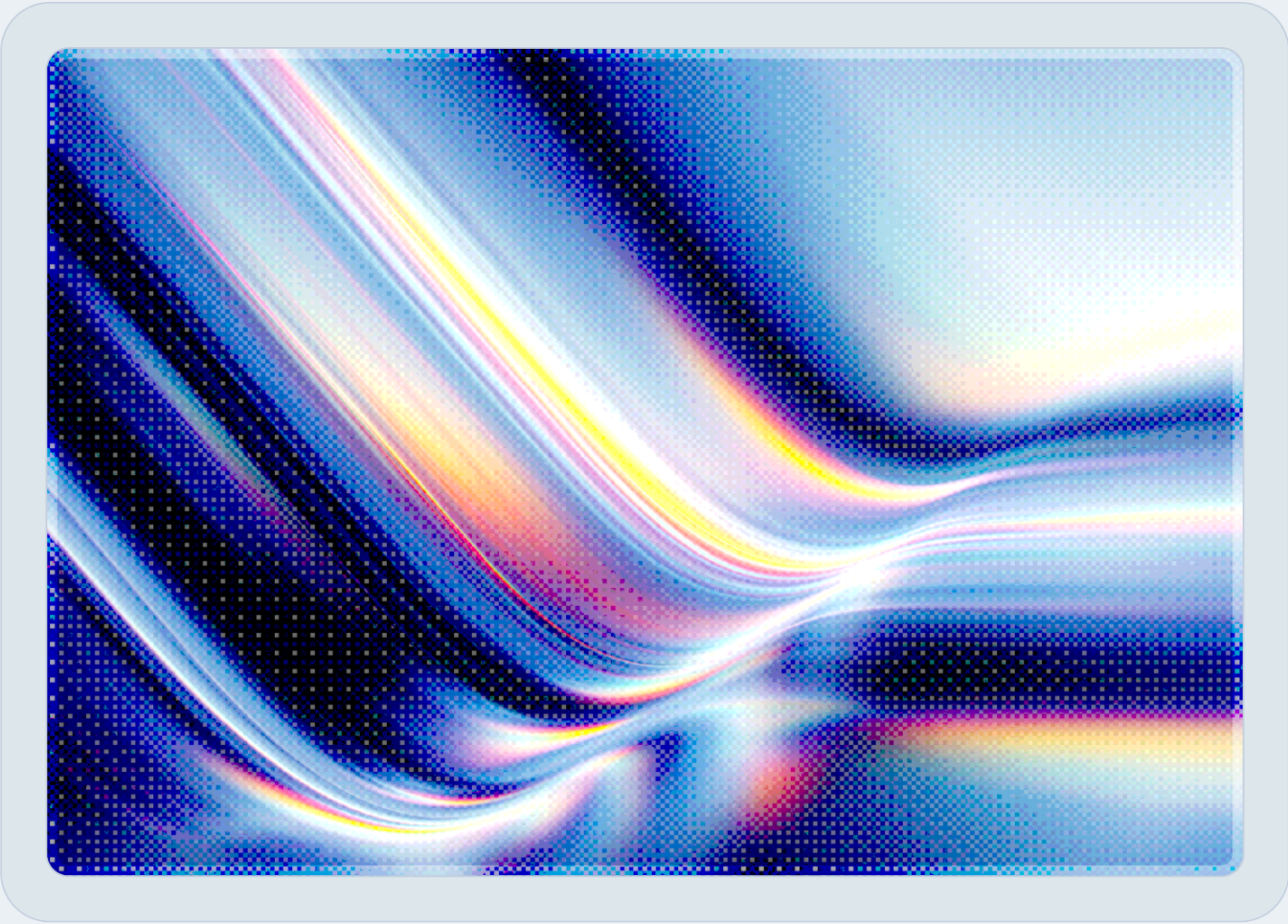
CSAI Foundation | Cloud Security Alliance

Sub-24-Hour Exploitation of AI Inference Frameworks

Rapid Attack Patterns Against Unauthenticated AI Serving Infrastructure and Security Guidance

2026-05-19

AI-assisted Rapid Research



CSAI

cloud security alliance®

Point-in-time analysis as of 2026-05-19. The CVE landscape covered in this document is actively evolving; patch status and exposure details should be verified against current vendor advisories.

Key Takeaways

- AI inference frameworks—including vLLM, Ollama, NVIDIA Triton, Meta Llama serving infrastructure, and NVIDIA TensorRT-LLM—have accumulated a critical mass of high-severity vulnerabilities in 2025 and 2026, many of which allow unauthenticated remote code execution with no prior access to the targeted system.
- A pattern of unsafe code reuse, collectively designated ShadowMQ by Oligo Security researchers, propagated an insecure use of Python's pickle deserialization through ZeroMQ-based inter-process communication across at least six major AI inference frameworks, producing multiple critical vulnerabilities in each. Thousands of exposed ZeroMQ sockets were identified on the public internet, directly tied to affected inference deployments. [1]
- CVE-2026-22778 (CVSS 9.8), disclosed in February 2026, enables unauthenticated remote code execution against vLLM deployments supporting multimodal inputs by chaining an information disclosure flaw with a heap buffer overflow in a bundled video processing dependency. The exploit requires only the ability to reach the API endpoint. [2]
- AI inference frameworks are frequently deployed without authentication by default, particularly in private cloud and on-premises GPU clusters, where operators assume network isolation provides adequate protection. Adversaries are actively disproving this assumption: honeypot infrastructure recorded more than 91,000 attack sessions targeting exposed LLM endpoints between October 2025 and January 2026, according to GreyNoise data published by Indusface. [3]
- The time between public vulnerability disclosure and active exploitation has compressed to hours for high-profile vulnerabilities, a trend confirmed by Synack's 2025 vulnerability research. [4] AI inference deployments are frequently managed on slower patch cycles than externally facing services, compounding exposure to this accelerating timeline.
- Security teams should treat AI serving infrastructure with the same urgency and rigor applied to externally facing web services, enforce authentication at all API boundaries regardless of network placement, and maintain a patch cadence matched to the speed at which vulnerabilities in this ecosystem are now being weaponized.

Background

The proliferation of self-hosted AI inference infrastructure is a relatively recent phenomenon, but it has already produced an attack surface whose breadth and exposure levels echo patterns from the early era of cloud computing—when organizations rushed to deploy services faster than security controls could be standardized. Frameworks such as vLLM, Ollama, NVIDIA's Triton Inference Server, Meta's Llama serving stack, and SGLang were designed primarily for performance and flexibility, enabling organizations to host large language models at scale on GPU clusters. Security hardening was not consistently integrated into the initial design of these frameworks.

The consequence is a class of infrastructure that is simultaneously high-value—sitting adjacent to proprietary model weights, sensitive inference data, and privileged compute environments—and weakly defended by default. Ollama, one of the most widely deployed local inference platforms, ships without authentication or access control mechanisms in its default configuration. [5] Triton Inference Server's HTTP and gRPC APIs are similarly unauthenticated unless explicitly secured. [9] The shared assumption underlying these defaults is that the inference layer will be isolated behind a trusted network perimeter, an assumption that is frequently violated in practice through misconfiguration, development environment exposure, or cloud network policy gaps.

Into this environment, 2025 and early 2026 introduced a sustained series of critical vulnerability disclosures. Pwn2Own Berlin 2026, held May 14–16, included AI inference systems among its competition categories, with researchers demonstrating successful exploitation across multiple AI frameworks. [6] The event awarded nearly $1.3 million for 47 zero-day vulnerabilities across all targets, and the research community's sustained focus on AI inference systems it reflects has placed this attack surface on the active adversarial radar in a way that was not previously true.

Published CVEs reached 48,244 during 2025, a 20 percent increase over 2024. [4] AI-related vulnerabilities with API-level exposure appear to represent a growing share of this volume, reflecting the rapid expansion of the inference infrastructure attack surface and the research community's intensifying focus on it.

Security Analysis

The ShadowMQ Pattern: How Code Reuse Became Systemic Risk

The ShadowMQ vulnerability class, disclosed by Oligo Security in November 2025, illustrates a risk pattern that is structurally distinct from typical software bugs: rather than an isolated coding error in a single product, it represents a shared implementation mistake that spread across multiple frameworks through code reuse. [1][13] The root cause is the use of ZeroMQ's `recv_pyobj()` method, which deserializes incoming network data using Python's `pickle` module. When this method is exposed over a network interface—a common pattern in distributed inference systems that use ZeroMQ for inter-process communication between worker nodes—an attacker who can reach the socket can transmit a crafted pickle object that executes arbitrary code on the host system.

The vulnerability affects Meta Llama's serving infrastructure, vLLM, NVIDIA TensorRT-LLM, Microsoft's Sarathi-Serve, Modular's Max Server, and SGLang—a substantial portion of the enterprise AI inference ecosystem in a single disclosure. [7] Specific CVEs include CVE-2025-30165 in vLLM (CVSS 8.0), CVE-2025-23254 in NVIDIA TensorRT-LLM (CVSS 8.8), and CVE-2025-60455 in Modular Max Server. [1] The fact that Oligo researchers identified thousands of exposed ZeroMQ sockets on the public internet, some explicitly associated with affected inference clusters, indicates that real-world deployments were directly at risk during the period between discovery and patching.

The ShadowMQ pattern carries an important lesson beyond the immediate remediation task. Inference frameworks have often been developed with performance and capability as primary design criteria, with security hardening addressed through subsequent iteration. Code reuse from convenience—importing a ZeroMQ-based communication pattern from another open-source project without auditing its security properties—can introduce critical vulnerabilities at scale. Organizations that evaluate AI inference frameworks before deployment should review not only the framework's own security posture but also the security properties of its inter-process communication layer and any bundled third-party components.

The CVSS 9.8 Scenario: Chained Exploitation in vLLM

CVE-2026-22778 (CVSS 9.8), disclosed on February 2, 2026, describes a pre-authenticated remote code execution vulnerability in vLLM versions 0.8.3 through 0.14.0 that affects deployments running multimodal inference endpoints. [2] The vulnerability represents a chained exploit, meaning it combines two distinct flaws to achieve an outcome that neither flaw could produce alone, and it illustrates how inference frameworks that integrate third-party media processing libraries inherit those libraries' attack surface.

The first stage exploits an information disclosure weakness in vLLM's error handling. When a multimodal endpoint receives an invalid image, the PIL image processing library throws an exception containing internal heap address information, which vLLM surfaces to the caller in its error response. An attacker submits a crafted invalid image, receives the error response, and extracts the leaked heap address from it. This breaks ASLR—the memory randomization protection that would otherwise make heap-based exploitation unreliable. The second stage uses this address to reliably exploit a heap buffer overflow in the JPEG2000 decoder bundled with the framework, achieving arbitrary code execution on the host system without requiring any credentials. [8]

The blast radius of a successful exploit against a vLLM deployment extends beyond the immediate server. vLLM is commonly deployed in clustered GPU environments where multiple worker nodes share a network segment, meaning an initial compromise of one inference endpoint can serve as a beachhead for lateral movement to adjacent GPU workers, the model weight storage layer, or other systems in the AI compute environment. Remediation requires upgrading to vLLM version 0.14.1; organizations that have not yet patched remain directly exposed to a publicly documented and technically detailed exploit chain.

Triton's Chain: From Information Leak to Full Server Control

NVIDIA's Triton Inference Server, a widely deployed inference platform in enterprise GPU environments, was the subject of a vulnerability chain disclosed by Wiz Research and patched in NVIDIA's August 2025 security bulletin. [9][14] The chain consists of three CVEs—CVE-2025-23319, CVE-2025-23320, and CVE-2025-23334—that individually produce limited impact but when combined allow an unauthenticated remote attacker to achieve full server control.

CVE-2025-23320 forms the entry point: an attacker sends an oversized request to the server, which triggers an internal error containing the name of a private shared memory region in the verbose error response. This name functions as a secret key for accessing the region. CVE-2025-23319 is an out-of-bounds write in the Python backend that fails to distinguish between user-owned and internally reserved shared memory keys; by registering the leaked key through the public API, an attacker gains read and write access to the server's private backend memory. CVE-2025-23334 is a separate out-of-bounds read that, combined with the write access from CVE-2025-23319, enables corruption of internal data structures sufficient to redirect execution. [9][12] The full chain requires no authentication at any stage.

The consequences of successful exploitation of a Triton server include theft of proprietary model weights—which may represent months of training compute and significant intellectual property investment—interception and manipulation of inference data being processed in real time, and use of the compromised server as a pivot point into the surrounding network. NVIDIA's patch, requiring upgrade to Triton version 25.07, addresses all three CVEs and should be treated as a mandatory remediation for any production Triton deployment. [9][14]

Exploitation Timeline Compression

The significance of these vulnerabilities is amplified by the speed at which threat actors now move from disclosure to active exploitation. Synack's analysis of vulnerabilities tracked across 2025 found that the exploitation window has compressed to hours for high-profile vulnerabilities—what previously required days or weeks of adversary preparation can now happen within a single business day. [4] AI-assisted reconnaissance and exploit development tools have been documented compressing the weaponization cycle, enabling adversaries to reach production exploitation within hours of a detailed technical writeup becoming public. [10]

This timeline compression creates a particularly dangerous environment for AI inference infrastructure. AI inference deployments are frequently treated as internal research infrastructure, and may therefore receive less frequent patching than externally facing services—a posture that the exploitation data from 2025 does not support. [4] Organizations may not yet have incorporated their GPU clusters and inference endpoints into their standard vulnerability management programs, or may treat them as internal research infrastructure rather than production systems requiring urgent patching. Data documenting over 91,000 attack sessions against LLM endpoints across a four-month window demonstrates that inference infrastructure is actively being probed and attacked at scale. [3] An unpatched vLLM or Triton deployment is not a low-priority asset waiting for a convenient maintenance window; it is a live target.

Recommendations

Immediate Actions

Organizations running vLLM versions 0.8.3 through 0.14.0 with multimodal endpoint support should treat CVE-2026-22778 as a critical remediation requiring immediate patching to version 0.14.1. The exploit chain is technically documented and pre-authentication; organizations running internet-accessible or multi-tenant deployments on affected versions should treat this as a critical remediation with no safe delay. NVIDIA Triton Inference Server deployments should be upgraded to version 25.07 to address the three-CVE chain. [9][14] Organizations running NVIDIA TensorRT-LLM, Meta Llama serving infrastructure, Modular Max Server, vLLM, or SGLang should audit for exposure to the ShadowMQ class of vulnerabilities and apply available patches; the Oligo Security advisory provides vendor-specific remediation guidance for each affected framework. [1]

Any Ollama deployment should be immediately audited for internet exposure. Because Ollama ships without authentication by default, a single misconfigured firewall rule or cloud security group can expose the full inference API to the public internet. The CVE-2025-51471 cross-domain token exposure vulnerability, disclosed in July 2025, demonstrates that even internal Ollama deployments handling model registry authentication can expose credential material to adversaries who can influence the model pull workflow. [11]

Short-Term Mitigations

Authentication should be enforced at the API gateway layer for all AI inference endpoints, regardless of whether the framework supports native authentication. A reverse proxy or API gateway configured to require token-based authentication provides a meaningful defense-in-depth layer for network-exposed vulnerabilities, requiring an attacker to obtain valid credentials before reaching the vulnerable service. Network segmentation should isolate inference clusters from general-purpose infrastructure, limiting the lateral movement opportunity available to an attacker who does achieve initial access.

ZeroMQ sockets used for inter-process communication within inference deployments should not be exposed on interfaces accessible from outside the inference cluster. Firewall rules should block external access to all ports used by ZeroMQ inter-process communication in the deployment; administrators should audit their configuration to identify these ports rather than relying on assumed defaults. Organizations should also audit their inference deployments for exposed ZeroMQ sockets using network scanning, as the thousands of publicly exposed sockets documented by Oligo suggest that many organizations are unaware of this exposure. [1]

Multimodal inference endpoints—those that accept image, audio, or video inputs—should be treated as a higher-risk attack surface than text-only endpoints, because they introduce media processing libraries (image decoders, video parsers, audio codecs) that carry their own vulnerability histories and are frequently less well-maintained than the inference framework itself. Network-accessible multimodal endpoints should be placed behind additional scrutiny, including input validation at the gateway layer and monitoring for anomalous error rates that may indicate active probing.

Strategic Considerations

AI inference infrastructure should be incorporated into standard enterprise vulnerability management programs with the same patching SLAs applied to externally facing web services. The evidence from 2025 and 2026 establishes that the distinction between "internal AI research infrastructure" and "production system requiring urgent security attention" is no longer tenable. Security teams should

establish asset inventory coverage for inference servers, GPU workers, model weight storage, and associated inter-process communication endpoints—all of which represent components of the AI compute attack surface.

Vendor security posture should be a factor in framework selection decisions. Organizations evaluating AI inference frameworks should assess the vendor's track record for vulnerability disclosure, patch responsiveness, and default security configuration, alongside the more commonly evaluated criteria of performance and model compatibility. A framework that ships with unauthenticated APIs by default, or that has a history of slow responses to reported vulnerabilities, introduces security debt that will compound over time. CSA's AI Controls Matrix (AICM) provides guidance on AI supply chain security controls applicable to framework selection and ongoing vendor management. [16]

Threat intelligence feeds covering AI-specific vulnerabilities should be subscribed to, as the pace of AI inference framework CVE disclosures in 2025 and 2026 may exceed the coverage depth of general-purpose vulnerability intelligence programs. The ShadowMQ pattern, the vLLM CVSS 9.8 chain, and the Triton vulnerability chain all received detailed public technical writeups before patches were universally applied—organizations with AI-focused threat intelligence could prioritize patching before active exploitation occurred, while those without dedicated coverage may receive insufficient early warning to patch before exploitation begins.

CSA Resource Alignment

This research note connects to several areas of CSA's published guidance on AI security. The MAESTRO framework for agentic AI threat modeling identifies the inference layer as a critical architectural tier, recognizing that compromise of inference infrastructure can affect not only the immediate AI application but also any downstream agentic systems consuming its outputs. [15] The attack patterns documented here—unauthenticated API access, model weight theft, and inference response manipulation—align with the threat categories MAESTRO addresses at the AI execution layer.

The CSA AI Controls Matrix (AICM) v1.0 provides directly applicable controls across several of its 18 domains. [16] The AI Supply Chain domain addresses the risks of inheriting vulnerabilities from third-party inference frameworks, which the ShadowMQ pattern exemplifies at scale. The AI Infrastructure Security domain maps to the network isolation and authentication controls recommended above. The AI Model Security domain addresses the risk of model weight theft, which is among the documented consequences of successful exploitation of Triton and vLLM. Organizations implementing AICM controls should treat the findings in this note as a prioritization signal for controls in these domains.

CSA's Zero Trust guidance is particularly relevant to the deployment posture issue at the center of this research note. The assumption that network placement—being "inside" a private cloud or on-premises cluster—provides adequate security for AI inference endpoints is exactly the kind of implicit trust that Zero Trust architecture is designed to eliminate. Applying Zero Trust principles to AI inference infrastructure means requiring explicit authentication and authorization for every API call to inference endpoints, regardless of whether the caller is internal or external to the network perimeter.

The CSA Cloud Controls Matrix (CCM) provides infrastructure-level control guidance applicable to the hosting environment for AI inference deployments, including controls for identity and access management, network security, and vulnerability management that address the foundational gaps documented in this research note. [17]

References

[1] Oligo Security. "ShadowMQ: How Code Reuse Spread Critical Vulnerabilities Across the AI Ecosystem." Oligo Security Blog, November 2025.

[2] OX Security. "Millions of AI Servers at Risk: Critical vLLM RCE Lets Attackers Take Over via Video Link (CVE-2026-22778)." OX Security Blog, February 2026.

[3] Indusface. "Exposed LLM Infrastructure: Risks and Exploits." Indusface Blog, April 2026.

[4] Help Net Security. "AI Shrinks Vulnerability Exploitation Window to Hours." Help Net Security, May 18, 2026.

[5] Dark Reading. "Ollama, Nvidia Flaws Put AI Infrastructure at Risk." Dark Reading, 2025.

[6] Infosecurity Magazine. "Security Researchers Find 47 Zero-Days at Pwn2Own Berlin." Infosecurity Magazine, May 2026.

[7] The Hacker News. "Researchers Find Serious AI Bugs Exposing Meta, Nvidia, and Microsoft Inference Frameworks." The Hacker News, November 2025.

[8] Orca Security. "Critical RCE in vLLM Allows Server Takeover via Malicious Video URL (CVE-2026-22778)." Orca Security Blog, February 2026.

[9] Wiz Research. "Breaking NVIDIA Triton: CVE-2025-23319 – A Vulnerability Chain Leading to AI Server Takeover." Wiz Blog, August 2025.

[10] Infosecurity Magazine. "AI-Enabled Adversaries Compress Time-to-Exploit." Infosecurity Magazine, March 2026.

[11] GitHub Advisory Database. "Ollama Vulnerable to Cross-Domain Token Exposure – CVE-2025-51471." GitHub, July 2025.

[12] NIST NVD. "CVE-2025-23319 Detail." National Vulnerability Database, 2025.

[13] CSO Online. "Copy-Paste Vulnerability Hits AI Inference Frameworks at Meta, Nvidia, and Microsoft." CSO Online, November 2025.

[14] NVIDIA. "Security Bulletin: NVIDIA Triton Inference Server – August 2025." NVIDIA, August 2025.

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

[16] Cloud Security Alliance. "AI Controls Matrix." Cloud Security Alliance, 2025.

[17] Cloud Security Alliance. "Cloud Controls Matrix." Cloud Security Alliance.