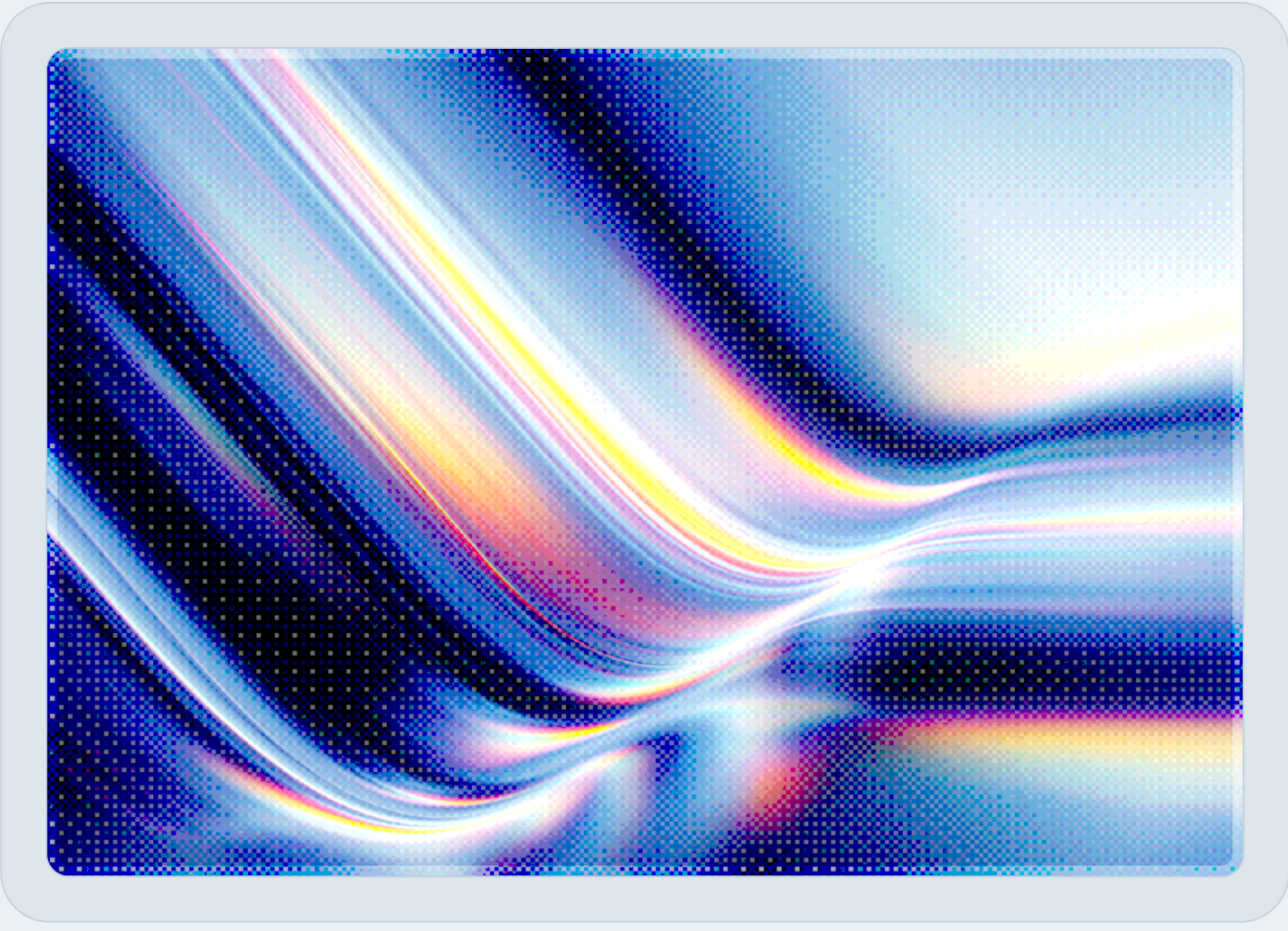
CSAI Foundation | Cloud Security Alliance

Confidential AI Under Pressure: SGLang RCE and SEV-SNP Attacks

Security Implications and Guidance from the NEAR AI Breach and the Mior and Skyfall Disclosures

2026-11-16

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

Two sets of events disclosed in the second week of November 2026 test the claims made for confidential AI inference from opposite directions. On November 9, NEAR AI reported that an attacker had used a remote code execution (RCE) vulnerability in SGLang, the open-source inference engine running on its servers, to gain code execution on an inference server and move into other parts of its infrastructure. The company acknowledged that requests processed on the compromised server, including confidential-mode requests, could have been observed [1][2]. Within days, AMD published advisories for two attacks on its SEV-SNP confidential virtual machine technology. Mior (CVE-2026-0462) lets a malicious host with hypervisor and BIOS control read and write guest memory and forge attestation reports. Skyfall (CVE-2026-40671) leaks guest register state through a transient-execution side channel [3][4][5].

The first lesson is that a trusted execution environment (TEE) does not protect a workload from its own software. Hardware isolation shields data from the host and operator, but it does nothing against an exploitable inference engine running inside the confidential VM, as NEAR AI's own report states [1]. The inference engine therefore belongs in the trusted computing base and should be treated that way. CERT/CC documented six SGLang vulnerabilities in July 2026, including unauthenticated RCE through pickle deserialization, with no patches available when its note was published [7], and public vulnerability trackers still listed no fixed release as of November 16, 2026 [13]. NEAR AI has not said which SGLang flaw was exploited in its incident, and no public source has linked the breach to these CVEs [1][2].

The outside-in protection also has gaps until platforms are patched. Mior and Skyfall both assume a malicious hypervisor, which is the adversary SEV-SNP is designed to stop, and full mitigation depends on OEM BIOS updates that AMD targets for November and December 2026 [3][4]. Attestation verification should therefore enforce minimum firmware versions. AMD's Mior advisory lists the `TCB[SNP]` values that accompany the mitigated SEV firmware; verifiers and key-release services that accept older values may trust an unpatched platform, and on an unpatched platform Mior can forge the attestation report [3][5]. Because the advisory lists these values alongside the SEV firmware fix rather than the separate BIOS-level fix, a TCB floor alone may not confirm that a platform is fully remediated. Finally, GPU confidential computing likely inherits these weaknesses. NVIDIA's reference architecture builds on an SEV-SNP or Intel TDX confidential VM [6], so a compromise of the CPU TEE would plausibly undermine GPU-backed confidential inference as well, although no such attack has been demonstrated.

Background

Confidential AI inference has become a selling point for providers that want to offer privacy guarantees stronger than contractual promises. In the typical architecture, the inference server runs inside a confidential virtual machine (CVM) protected by AMD SEV-SNP or Intel TDX. The CVM is paired with GPUs operating in confidential computing mode, and remote attestation lets a client or key broker verify the platform and software before sending prompts or releasing model keys. NVIDIA's reference implementation for deploying proprietary models on confidential compute follows this pattern. It supports SEV-SNP on EPYC 7003 and 9004 processors or TDX on recent Xeon processors, and model decryption keys are released only after both CPU and GPU evidence has been verified against an approved policy [6]. The resulting promise to users is that neither the cloud operator nor the inference provider can read their prompts and completions.

NEAR AI operates NEAR AI Cloud, a TEE-backed confidential inference service, and previously offered a consumer Private Chat product that it had deprecated before the incident and took offline after detection [1]. In October 2026 it announced that Brave's Leo assistant would use its infrastructure, although the incident report does not identify Brave users among those affected [2]. According to NEAR AI's November 9 incident report, an attacker exploited an RCE vulnerability in SGLang to execute code on a single inference server and used that foothold to pivot further into the company's infrastructure [1]. The exposed data included names and email addresses for about 100 users, with plan and balance information for some, along with chat histories, settings, and logs for 18 agents on a legacy deployment that did not run in the TEE. API keys belonging to two NEAR AI Cloud customers had also been written to logs and were revoked [1][2]. NEAR AI reported no indication that stored Private Chat conversations or uploaded files were accessed. It stated that requests processed on the compromised server on November 8, including confidential-mode requests, could have been observed, but it has not said whether any were actually read [1].

NEAR AI did not identify which SGLang vulnerability was exploited, and media coverage has not linked it to a specific CVE [2]. The broader state of SGLang security is still relevant context. CERT/CC's vulnerability note VU#281278, published July 30, 2026, describes six SGLang flaws. They include unauthenticated RCE through pickle deserialization on the `/load_lora_adapter_from_tensors` endpoint (CVE-2026-15969), server-side request forgery and local file read through `image_url` in `/v1/chat/completions` (CVE-2026-15974), RCE through a `torch.load` fallback on `/update_weights_from_disk` (CVE-2026-15976), API key leakage through `/server_info` (CVE-2026-15977), and model weight exfiltration via NCCL broadcasting (CVE-2026-15978). CERT/CC reported that coordination with maintainers was unsuccessful and no patches were available at publication [7]. As of November 16, 2026, public

vulnerability tracking for CVE-2026-15969 still reported no fixed release [13]. NEAR AI states that it deployed its own fix and reported the flaw it encountered upstream [1], but whether the intrusion used one of the CERT/CC flaws or a different one has not been publicly established.

The AMD disclosures arrived in parallel. AMD's bulletins AMD-SB-3059 (Mior) and AMD-SB-3061 (Skyfall) both carry an initial publication date of November 10, 2026, and the research was publicly reported on November 13 [3][4][5]. Mior comes from ETH Zurich researchers [8]. Skyfall comes from researchers at ETH Zurich and Google and builds on the Transient Scheduler Attacks class first disclosed in July 2025 [5].

Security Analysis

Two Threat Models, One Trust Boundary

The two incidents are attacks against the two sides of the confidential computing trust boundary. SEV-SNP and TDX are designed to defend against an outside-in adversary: a compromised or malicious hypervisor, host operating system, or cloud operator trying to read or tamper with guest memory. They are not designed to defend against an inside-out adversary who compromises software running inside the guest. Such an attacker has the same view of plaintext data that the legitimate workload has. NEAR AI's report makes this distinction explicit, stating that confidential computing protects data from the operator and the host but not from vulnerabilities in software running inside the environment [1].

The table below summarizes how each event relates to that boundary, including who can act on remediation and what a tenant can do on its own.

Dimension	SGLang RCE at NEAR AI	Mior (CVE-2026-0462)	Skyfall (CVE-2026-40671)
Attacker position	Remote, via network-reachable inference software	Malicious host with hypervisor and BIOS/UEFI control plus a PCIe device	Malicious host or hypervisor; no hardware tampering
Direction	Inside-out (compromise of software in the TCB)	Outside-in (breaks SEV-SNP integrity)	Outside-in (breaks SEV-SNP confidentiality)

Dimension	SGLang RCE at NEAR AI	Mior (CVE-2026-0462)	Skyfall (CVE-2026-40671)
Effect	Code execution where confidential requests are processed; lateral movement	Arbitrary guest memory read/write; forged attestation reports	Leakage of guest register state and limited cache lines
Severity indicator	Not published for the specific flaw	CVSS 4.0: 5.9 [3]	CVSS 4.0: 4.9 [4]
Remediation owner	Inference provider (software patching, architecture)	Cloud provider / OEM (SEV firmware and BIOS)	Cloud provider / OEM (microcode and BIOS)
Tenant can self-remediate?	Yes, if self-hosting SGLang (restrict endpoints, apply fixes when available); otherwise depends on provider	No; tenant can only verify TCB through attestation	No; tenant can only verify TCB through attestation

The Inference Engine Is Part of the Trusted Computing Base

The NEAR AI incident shows that a confidential inference system's trusted computing base (TCB) includes more than the CPU, firmware, and guest kernel. It includes the serving engine, its Python dependency tree, tokenizer and multimodal preprocessing code, model and adapter loading paths, and any administrative endpoints exposed on the same network interface. Inference engines are often large, fast-moving codebases optimized for throughput, and several of the SGLang flaws in the CERT/CC note sit in features such as dynamic LoRA adapter loading and weight updates [7]. These features deserialize attacker-influenced data, and pickle deserialization in particular can lead directly to code execution, as CVE-2026-15969 illustrates [7]. Remote attestation can show that a particular build of SGLang is running. It cannot show that the build is free of exploitable bugs.

The incident also suggests that the blast radius of an inside-out compromise depends heavily on what else the TEE can reach. According to the report, the attacker pivoted from a single inference server into other infrastructure and reached an internal system connected to the deprecated Private Chat feature, and customer API keys turned up in logs [1][2]. That pattern suggests the confidential workload held credentials or network paths that reached beyond its own confidential VM, although the public report

does not detail the lateral movement path. Flaws such as the SGLang SSRF (CVE-2026-15974), which can expose internal metadata and secrets [7], illustrate why egress from inference nodes matters even where it is not known to have been used. A design in which each confidential inference node holds only the minimum credentials, cannot initiate connections to unrelated internal services, and never logs request-bearing secrets would be expected to limit the damage of a similar compromise. NEAR AI has said it will apply the same security rigor to the inference software and surrounding services as to the enclave itself [1].

The exposure of 18 agents on a legacy deployment that ran outside the TEE points to a related governance issue. Confidential and non-confidential deployments can coexist within a single provider, and customers may not always know which of their workloads run where. Providers making confidentiality claims should be precise about which product surfaces those claims cover.

Mior: When the Host Can Rewrite the Rules

SEV-SNP's integrity guarantees depend on the Reverse Map Table (RMP), a structure that records which guest owns each page of physical memory and that the hypervisor is not supposed to be able to modify arbitrarily. According to the researchers' description, Mior manipulates memory-mapped I/O routing during SEV-SNP initialization. The Platform Security Processor's writes to initialize the RMP are redirected to an attacker-controlled PCIe device, a GPU in the demonstration, instead of main memory [5][8]. Initialization completes without errors, but part of the RMP, including the entry protecting the table itself, remains uninitialized. The hypervisor can then rewrite RMP protections, enable debugging on a confidential VM that had disabled it, and use firmware debug commands to read and write guest memory. The researchers reported success in all 1,000 attempts across two test runs on an EPYC 9135 (Zen 5) processor, with the optional ciphertext hiding feature enabled [5]. Zen 3 and Zen 4 systems were not tested because the team had no BIOS setup it could modify on those platforms [8].

For confidential AI, the most consequential capability is attestation forgery. In a second experiment, the researchers copied a benign VM's launch measurement into a malicious VM so that its signed attestation reports described the benign image [5]. Confidential inference architectures typically gate model key release and client trust on exactly this kind of report [6]. An attacker who controls an unpatched host could therefore present a convincing attestation for a VM that does not provide the advertised protections. The preconditions are demanding: control of the hypervisor and BIOS/UEFI, with a test BIOS modified to leave memory-routing settings unlocked [5]. These preconditions nonetheless fall within the malicious-operator threat model that SEV-SNP is marketed to address, since AMD's threat model treats UEFI as untrusted [8].

AMD rates CVE-2026-0462 at CVSS 4.0 base score 5.9 and lists SEV-SNP, but not SEV or SEV-ES, as affected across EPYC 7003, 8004, 8005, 9004, and 9005 processors [3]. Remediation has two components. SEV firmware mitigations were released on August 12, 2026, while the platform initialization (PI) firmware that OEMs must integrate into BIOS updates is targeted for December 2026 [3]. For verifiers, the advisory publishes the TCB values associated with the mitigated SEV firmware: `TCB[SNP]=0x1F` for EPYC 7003, `0x1E` for EPYC 8004 and 9004, and `0x8` for EPYC 8005 and 9005 [3]. The advisory lists these values next to the SEV firmware fix rather than the PI firmware fix, and reporting notes that it does not say how VM owners should check the latter [5]. Enforcing the TCB floor is therefore necessary but may not by itself demonstrate that the BIOS-level fix is installed. AMD's two advisories also give different target months, November in AMD-SB-3061 and December in AMD-SB-3059, for the same TurinPI package [3][4], so tenants should confirm actual BIOS deployment dates with their providers.

Skyfall: Side Channels Through Failed Launches

Skyfall takes a different route. AMD describes it as incomplete cleanup during VMRUN error handling that can expose SEV-SNP guest register state to a privileged attacker through a Transient Scheduler Attacks side channel [4]. As reported, when a hypervisor deliberately causes a VM launch to fail, part of the VM's saved state has already been cached and is not cleaned up, so the hypervisor can read it [5]. The researchers recovered saved register values and a disk encryption key from a test VM, pausing the kernel at a point where the key was held in registers. They noted that a real attacker might reach that point using single-stepping techniques such as SEV-Step. Memory leakage is limited to three fixed cache-line offsets within a targeted page [5]. The same work also recovered decrypted microcode updates and register values from System Management Mode memory, the latter tracked separately as CVE-2026-18448 under AMD-SB-7072 [5][9].

Skyfall was demonstrated on Zen 3 (EPYC 7313) and Zen 4 (EPYC 9124 and 9334) processors and was not shown on Zen 5 [5]. According to the IEEE Symposium on Security and Privacy program committee's review, as reported by The Hacker News, a realistic attack would require an otherwise idle core, rapid repeated failed VM launches with performance-counter reads, and a victim VM that does not crash or time out [5]. On remediation, microcode mitigations were released on August 12, 2026, platform firmware is targeted for November 2026, and embedded EPYC updates are expected between January and April 2027 [4][5]. AMD-SB-3061 also lists the Instinct MI300A accelerator APU as affected, with a patch dated September 21, 2026 [4], which is directly relevant to AI hardware. Inference servers hold long-lived decryption and session keys, however, and because Skyfall recovered a disk encryption key from registers in testing, even narrow register leaks deserve attention in that setting.

Implications for GPU Confidential Computing

GPU confidential computing does not stand on its own. NVIDIA's reference architecture requires an SEV-SNP or TDX confidential VM as the host-side TEE and relies on combined CPU and GPU attestation before releasing keys [6]. The GPU's own attestation establishes that a genuine, non-revoked GPU is running approved firmware, but data reaches the GPU through the CVM. If the CVM's memory integrity or attestation can be subverted, the protection of prompts, completions, and model weights in transit to and from the GPU may be undermined regardless of the GPU's own state. This is an inference from the architecture rather than a demonstrated attack. It nonetheless means that organizations evaluating confidential AI offerings on AMD platforms should treat Mior and Skyfall remediation status as directly relevant.

Recommendations

Immediate Actions

Organizations running SGLang should immediately inventory every deployment and confirm that no SGLang HTTP or IPC interface is reachable from untrusted networks. CERT/CC recommends restricting network access, segmenting inference hosts, disabling unused endpoints, and setting `SGLANG_USE_PICKLE_IPC` to `false` [7]. CERT/CC's vulnerability descriptions also indicate that the weight-broadcasting flaw (CVE-2026-15978) applies when no API keys are configured, and that `/server_info` leaks keys when only `--admin-api-key` is set (CVE-2026-15977) [7]. Teams should therefore configure both a user API key and an admin key, and should block or disable `/server_info` at the proxy layer until upstream fixes are available. Because no fixed release had been reported as of November 16, 2026 [13], these compensating controls are the primary defense, and teams should track upstream fixes for the six CVEs closely. Customers of NEAR AI Cloud should confirm whether they were notified, rotate any API keys that may have appeared in logs, and review account activity, consistent with NEAR AI's own guidance to watch for unsolicited messages about accounts or balances [1][2].

Consumers of SEV-SNP-based confidential computing, whether directly or through a confidential AI service, should ask their providers for the remediation status of AMD-SB-3059, AMD-SB-3061, and AMD-SB-7072. Remediation must come from the platform owner and OEM, and tenants can only verify it [3][4][9]. Owners of attestation verifiers and key-release policies should update minimum TCB requirements to the mitigated `TCB[SNP]` values published in AMD-SB-3059 and reject reports below them once their providers have deployed the fixes [3]. Because those values correspond to the SEV

firmware mitigation, tenants should also obtain written confirmation from their providers that the BIOS/PI firmware updates have been installed, since attestation alone may not reveal the PI firmware version.

Short-Term Mitigations

Over the next one to three months, confidential inference operators should treat the serving engine as hostile-input-facing code that sits inside the TCB. In practice, that means running the inference engine in a separate, least-privileged process or sandbox within the CVM. It also means removing administrative and weight-update endpoints from production images, and refusing to load adapters or weights from any path that accepts serialized Python objects. Operators should ensure that a compromised inference node holds no credentials that grant access beyond its own function and cannot initiate connections to unrelated internal systems. Logging pipelines should be reviewed so that API keys, prompts, and completions are never written to logs, given that customer keys were exposed through logs in the NEAR AI incident [1].

On the attestation side, key brokers should bind release decisions to the full chain of evidence: CPU TEE TCB version, guest launch measurement, GPU attestation and revocation status, and the specific inference software build. NVIDIA's reference pattern of releasing model keys only after joint CPU and GPU evidence has been verified against an explicit policy is a reasonable baseline [6]. That policy now needs to include the post-Mior TCB floor, supplemented by provider confirmation of BIOS-level fixes. Operators should also publish clear documentation of which product surfaces run inside TEEs, so that legacy or auxiliary services are not mistaken for confidential ones.

Strategic Considerations

Confidential computing should be positioned as one layer of defense, not as a guarantee. Mior is the latest in a series of SEV-SNP integrity attacks, following earlier research such as RMPocalypse and Fabricked [5]. The pattern suggests that hardware TEE assurances will continue to be revised as researchers probe them. Procurement and risk teams should evaluate confidential AI offerings on demonstrated patch cadence, transparency about TCB versions, reproducible and publicly verifiable software builds, and incident disclosure practices, in addition to the hardware feature list. NEAR AI's publication of its confidential inference stack's code and configurations makes independent verification possible [1], although published code did not by itself prevent the incident. Providers making privacy claims to end users should describe the outside-in threat model those claims cover and acknowledge the inside-out risks they do not.

For the open-source inference ecosystem, the SGLang disclosure history argues for sustained investment in security review, coordinated disclosure capacity, and safer defaults. Moving away from pickle-based IPC, which CERT/CC notes the maintainers have begun, is one example, although the insecure setting remained the default when CERT/CC published [7].

CSA Resource Alignment

CSA's **Data Security within AI Environments** maps data security controls across the AI lifecycle and identifies gaps in the AI Controls Matrix for protecting data in AI systems [10]. The NEAR AI incident falls within its scope. Prompts and completions are data in use, and the incident shows that protecting them requires controls on inference software, logging, and credential handling in addition to hardware isolation.

CSA's Health Information Management Working Group paper **Healthcare Confidential Computing and the Trusted Execution Environment** explains how TEEs and confidential computing protect sensitive data in use, including for AI and machine learning workloads [11]. The Mior and Skyfall disclosures are a reminder that the assurances described there depend on current platform firmware and on attestation policies that enforce it. Healthcare and other regulated organizations adopting confidential AI should pair the paper's architectural guidance with the TCB verification practices described in this note.

The **AI Controls Matrix (AICM) v1.1** provides the control framework for operationalizing these recommendations [12]. Its Threat and Vulnerability Management domain covers tracking and remediating inference-engine CVEs such as those in VU#281278. Its Supply Chain Management, Transparency, and Accountability domain supports provider due diligence on TEE patch status and attestation transparency. Its Identity and Access Management and Logging and Monitoring domains address the credential scoping and log handling weaknesses illustrated by the NEAR AI incident.

References

[1] NEAR AI. "Security Incident Report: SGLang Vulnerability." NEAR AI Blog, November 9, 2026.

[2] Jose Antonio Lanz. "Near AI Breach Exposes User Details and Data From 18 AI Agents." Decrypt, November 10, 2026.

[3] AMD. "AMD Secure Encrypted Virtualization (SEV) Vulnerability, AMD-SB-3059 (CVE-2026-0462)." AMD Product Security, November 10, 2026.

[4] AMD. "AMD SEV Incomplete Cleanup Vulnerability, AMD-SB-3061 (CVE-2026-40671)." AMD Product Security, November 10, 2026.

[5] Swati Khandelwal. "New Mior and Skyfall Attacks Let Rogue Cloud Hosts Steal Secrets From AMD SEV-SNP VMs." The Hacker News, November 13, 2026.

[6] NVIDIA. "Reference Implementations: Deploying Proprietary Models Securely with NVIDIA Confidential Computing on Self-Hosted Virtual Machines." NVIDIA Enterprise Reference Architectures, accessed November 16, 2026.

[7] CERT/CC. "VU#281278: SGLang Contains Six Different Vulnerabilities Including RCE, Data Exfiltration, and Credential Disclosure." CERT Coordination Center, July 30, 2026.

[8] Chris Wech, Philipp Giersfeld, Benedict Schlüter, and Shweta Shinde. "Mior: Breaking AMD SEV-SNP with Device-as-Memory." Proceedings of ACM CCS '26, November 2026.

[9] AMD. "EDK II Mitigations for Transient Scheduler Attacks (TSA), AMD-SB-7072 (CVE-2026-18448)." AMD Product Security, November 10, 2026.

[10] Cloud Security Alliance. "Data Security within AI Environments." Cloud Security Alliance, 2025.

[11] Cloud Security Alliance. "Healthcare Confidential Computing and the Trusted Execution Environment." CSA Health Information Management Working Group, 2025.

[12] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2026.

[13] HOL Guard. "CVE-2026-15969." HOL Guard CVE Database, accessed November 16, 2026.

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