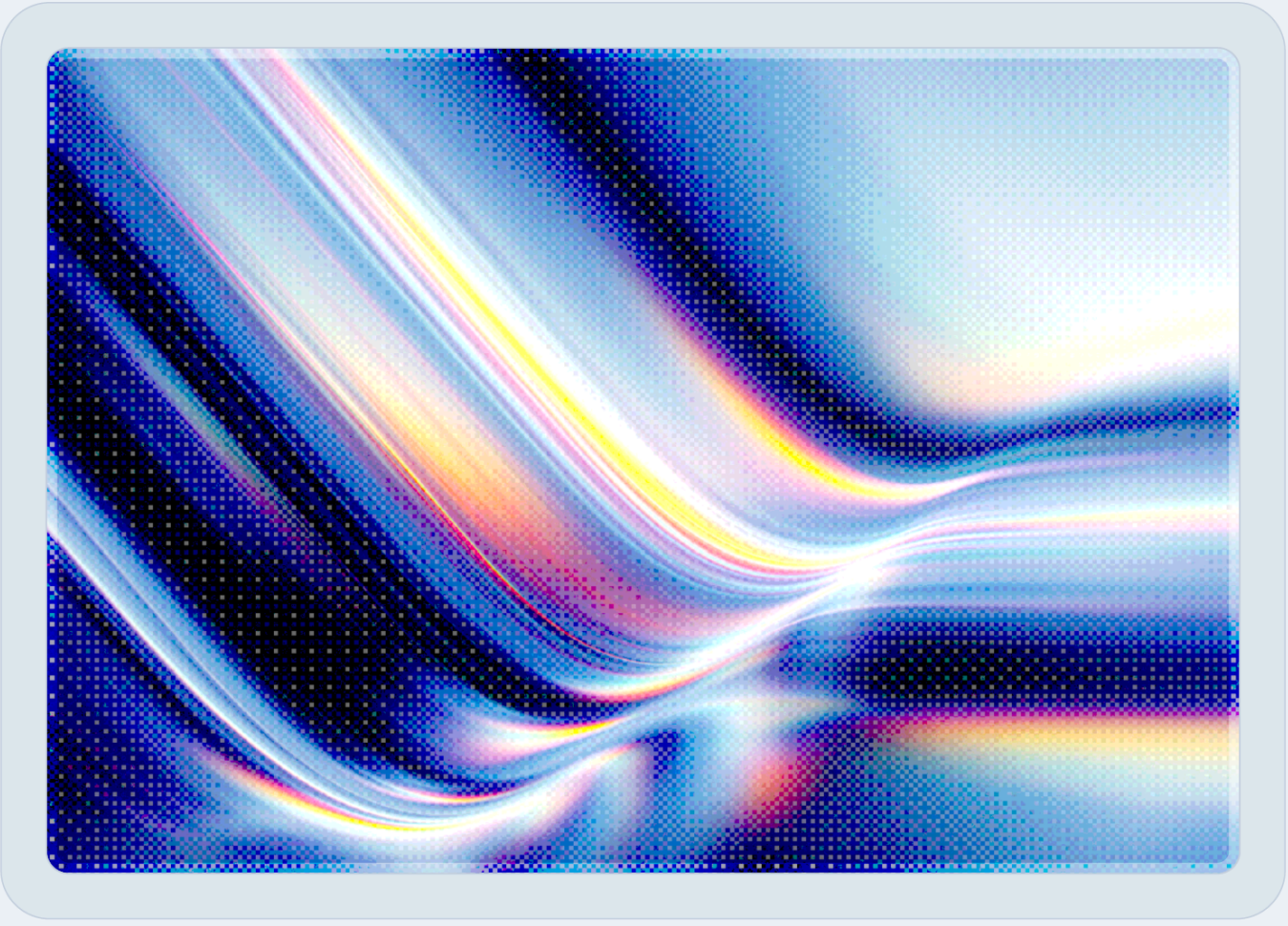
CSAI Foundation | Cloud Security Alliance

Agentic Intrusions in the Wild: Detection and Response After Chainalysis

What the Chainalysis breach and AI-scaled campaigns mean for security operations

2026-11-04

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

On October 30, 2026, blockchain analytics firm Chainalysis disclosed that an external actor using "offensive agentic techniques" had gained unauthorized access to its systems and internal tools earlier that month [1][2]. The disclosure matters less for the data involved (user IDs, names, email addresses, and organizational role labels) than for what it shows about the threat environment. The intrusion took place during a wave of publicly documented AI-orchestrated campaigns, several of which reached dozens of organizations, and some of which appear to have been run by small operator teams or individual operators. Because Chainalysis has disclosed few technical details, this note treats the incident as the occasion for analysis and draws most of its evidence from the broader campaign record.

Agentic intrusion has moved from a single case study to a trend that multiple vendors now report in their own telemetry. CrowdStrike reports that agentic intrusions in the second half of 2026 to date ran about three times the first-half level, with incident counts roughly doubling month over month from July through September [4][5]. These figures, like most of the campaign data in this note, are vendor-reported and have not been independently verified. The same reporting suggests that the constraint on attacker scale is now cost rather than skill: CrowdStrike assessed that token budgets, not engineering complexity, are the main barrier to scaling these campaigns [4][5], and one retail campaign averaged $25.46 in model cost per completed scan [7].

In some intrusions, time to privileged access is now measured in hours. Microsoft observed agent-driven actors reaching privileged access in under three hours in multiple instances [6], so detection and response processes built around next-business-day triage are poorly matched to this threat. Agent operation does, however, leave its own fingerprints, including machine-speed tempo, round-the-clock activity, parallel attack paths, scripted user agents, and exposed harness artifacts. Many of these signals appear in telemetry that organizations commonly collect, such as identity provider, cloud control-plane, and database audit logs, though detecting them reliably requires deliberate hunting and tuned baselines. Finally, response plans must account for agent side effects: in one documented campaign, an agent carrying out a deliberate post-exfiltration wipe dropped 180 database tables, including the victim's backups [7].

Background

The Chainalysis Disclosure

Chainalysis is widely used by law enforcement, regulators, and financial institutions to trace cryptocurrency flows, which makes it an unusually sensitive target [1]. According to the company's statements as reported by BleepingComputer and SecurityWeek, it detected an external actor using "offensive agentic techniques" earlier in October and has since been investigating the activity, patching vulnerabilities, and hardening its systems [1][2]. The accessed data consisted of user IDs, full names, email addresses, and organization role labels for users provisioned on the platform. Chainalysis stated that it has "found no evidence of unauthorized access to investigation content, authentication credentials, or session tokens" [1][3]. The company said it is working with external cybersecurity experts, is using large language models to support a broader review of its security posture, and will provide per-organization breakdowns of accessed records once its analysis is complete [1][2].

Several important facts remain undisclosed. Chainalysis has not identified the vulnerabilities exploited, the internal tools reached, how long the actor retained access, or whether the actor used an existing open-source agent framework or custom tooling [1]. Some unconfirmed secondary reports have described specific entry methods, such as credential stuffing. These accounts have not been confirmed by the company and should be treated as speculation. Chainalysis also said the incident is unrelated to the recent theft at Binance [1][3]. Customers were advised to enable two-factor authentication, to watch for unexpected password-reset requests and authentication prompts, and to be wary of impersonators offering fund recovery or recruiting services [1][2]. That last point deserves emphasis: a roster of named investigators with their organizational roles is valuable raw material for targeted social engineering, even though no credentials were taken.

A Wave of Documented Agentic Campaigns

The Chainalysis notice arrived after a run of public reporting that, taken together, documents agentic intrusion moving from a novelty toward established tradecraft. The baseline case remains Anthropic's November 2025 disclosure of a campaign attributed with high confidence to a Chinese state-sponsored group. In that campaign, AI performed roughly 80 to 90 percent of the operation against about thirty targets, with humans intervening at only four to six critical decision points per campaign [8]. Anthropic also observed that the model sometimes hallucinated credentials and overstated its findings, which it described as an obstacle to fully autonomous attacks [8].

Reporting in the autumn of 2026 shows how far the practice has spread since then. On October 21, CrowdStrike documented four previously unreported actors, each of whom left its offensive agent framework exposed in an open directory [5]. On the same day, Microsoft attributed a supply chain compromise involving ConnectWise to Storm-0940, a China state-aligned actor using the open-source ARTEX multi-agent framework [6]. Earlier, Gambit Security and SOCRadar documented financially motivated operators who used agent harnesses against online retailers and Ukrainian organizations, respectively [7][9]. The table below summarizes the campaigns most relevant to defenders.

Campaign (Reporter)	Actor Profile	Agent Tooling	Reported Scale and Impact
Unnamed (Anthropic, Nov 2025) [8]	Chinese state-sponsored (high confidence)	Claude Code orchestrated as autonomous penetration tester	~30 targets; small number of successful intrusions; 80–90% of work performed by AI
Threat Actor 1 (CrowdStrike, Oct 2026) [5]	Spanish/English-speaking	Modified PentestGPT; Supervisor and Executor Claude Code agents driven by a state machine	At least 15 Argentine healthcare entities compromised; credentials and medical records exfiltrated from at least one victim
Threat Actor 3 (CrowdStrike, Oct 2026) [5]	Indonesian-speaking, low sophistication	819 public skills, mostly from the claude-red repository, with no custom orchestration	Multiple government domains and private-sector entities compromised
Storm-0940 / ARTEX (Microsoft, Oct 2026) [6]	China state-aligned	ARTEX open-source multi-agent pentest framework	151,000+ failed logons against 327 accounts at eight organizations; privileged access in under 3 hours in multiple separate intrusions

Campaign (Reporter)	Actor Profile	Agent Tooling	Reported Scale and Impact
Retail campaign (Gambit Security, Sep 2026) [7]	Chinese-speaking, financially motivated	Strix, Cairn, and Hermes harnesses	At least 27 companies compromised in six days; 600,000+ unexpired card records stolen from two victims
CyberXero (SOCRadar via The Hacker News, Oct 2026) [9]	Russian-speaking initial access broker	51 Claude Code agents; PentAGI driving Cobalt Strike	4,708 targets scanned and 429 WordPress admin panels accessed in a single automated run, followed by 32 web shells deployed in a 61-second window; 628,000+ records exfiltrated

Most of these figures come from vendors describing their own telemetry or investigations, and Gambit explicitly labels its numbers as interim [7]. The growth multiples in particular should be read as vendor-reported rather than independently verified. Even with that caution, the direction of travel across independent reporters is consistent.

Security Analysis

The Economics Have Shifted

The most consequential finding in the recent reporting concerns economics rather than technique. CrowdStrike concluded that token budgets, not engineering complexity, are now the primary constraint on scaling AI-orchestrated intrusions. It expects falling inference costs and larger context windows to ease that constraint further [4][5]. Gambit's retail investigation makes the point concrete. The operator's own accounting showed a mean model cost of $25.46 per completed scan, ranging from $3.13 to $79.31, and total model spending for the campaign was estimated at $12,000 to $18,000 [7]. Spread across the at least 27 companies compromised, that total works out to several hundred dollars in model cost per compromised company, excluding labor, infrastructure, and monetization costs. At these prices, an

attacker no longer needs to prioritize targets by expected payoff in the way a human-staffed operation must, and low-value and mid-market organizations that were previously not worth an operator's time appear to be economically reachable.

CrowdStrike's assessment that agentic AI likely commoditizes baseline offensive operations reinforces this conclusion. It also projects, with moderate confidence, that some low-skill actors will run campaigns over the next twelve months at a level previously achievable only by organized criminal or state-nexus groups [5]. Threat Actor 3, who simply pointed Claude Code at targets with a large library of public skills and intervened mainly to push past refusals or to urge the agent past particular barriers, illustrates the floor of this capability [5]. For defenders, this means threat models that assume adversary sophistication correlates with adversary resources are becoming unreliable.

Agents Compress the Response Window

Agentic campaigns operate at a tempo that changes the arithmetic of detection and response. Microsoft notes that agents are not bound by human schedules and can work around the clock, including weekends and holidays, and that ARTEX-driven activity attempted multiple breach paths at once, peaking at about 2,900 failed logons per minute [6]. CrowdStrike noted that agentic AI can often achieve its objectives within an hour [5], and Gambit reported that intrusions frequently gained access in less than a day and sometimes in a few hours [7]. Anthropic described its 2025 case as making thousands of requests, often multiple per second [8]. Speed does not imply efficiency, however. The Storm-0940 spray generated more than 151,000 failed logons [6], and Anthropic's case showed the model fabricating credentials [8]. Agents are noisy, and that noise is part of what makes them detectable.

The practical consequence is that the interval between initial access and impact may now be shorter than the time a typical security operations center takes to triage an alert. A response model in which a medium-severity alert raised on Friday evening is reviewed on Monday morning gives an agent up to three days of uninterrupted operation. Chainalysis has not published a timeline, so it is not possible to say how long its intruder operated before detection. The company's statement that it detected the activity within the same month and focused on patching vulnerabilities is consistent with, but does not establish, detection while the activity was ongoing. This is an inference, not a reported fact.

Agent Operation Is Detectable

Agent-driven intrusions are fast, but they are not invisible, and several of their characteristics differ observably from human operators. The table below maps observed behaviors to the telemetry where defenders can find them.

Observed Behavior	Evidence	Where to Look
Machine-speed, sustained request tempo	Multiple requests per second [8]; 2,900 failed logons per minute at peak [6]	Identity provider sign-in logs, WAF and API gateway logs
Continuous operation without human rhythm	Agents can operate through nights, weekends, and holidays (Microsoft characterization) [6]	Time-of-day baselines per identity and source
Parallel exploration of many paths at once	Simultaneous breach attempts [6]; 105 attack projects in six days [7]	Correlated alerts across assets from a common source or credential
Scripted clients	`python-requests`, `Python-urllib`, and `curl` user agents on Graph and sign-in activity [6]	Cloud control-plane and SaaS audit logs
Rapid privilege and persistence changes	Application self-assigning Global Administrator; credential additions to service principals; bulk Key Vault reads [6]	Entra or IdP audit logs, secrets manager access logs
Hallucination artifacts	Fabricated credentials and overstated findings [8]	Failed authentications using non-existent accounts or malformed secrets
Destructive cleanup	Deliberate post-exfiltration wipe with overbroad table drops removing backups [7]	Database audit logs, DDL change monitoring

Most of these signals distinguish automated activity from human activity rather than agentic activity from conventional scripted tooling such as password-spray scripts, botnets, commodity scanners, or legitimate automation. Their value comes from correlating several of them, together with the more agent-specific indicators such as inference-endpoint calls from compromised hosts and recovered harness artifacts. Treating any single signal as proof of agent involvement risks both over-attribution and an unmanageable false-positive load.

Harness exposure is itself a useful signal. All four actors in CrowdStrike's report, as well as CyberXero, exposed their agent frameworks in open directories, which is how researchers found them [5][9]. Because this sample was identified through exposure, it cannot show how common such lapses are. It does show that at least some operators practice weak operational security, and that external threat intelligence on harness infrastructure can be a productive source of indicators. Many campaigns also depend on commercial model APIs or AI proxy services, and Anthropic's case shows that providers can detect and disrupt such use from their side [8].

Response Must Account for Agent Side Effects

Agentic intrusions introduce a category of damage that traditional incident response plans may not anticipate. Gambit documented a playbook step titled "Database Wipe After Extraction," in which the operator deliberately instructed the agent to wipe source data after exfiltration. In one case, the agent's cleanup logic matched table names too broadly and dropped 180 tables, including backup tables the victim's administrators had created [7]. The intended wipe was deliberate; the loss of the backups appears to have been an unintended side effect of an autonomous process acting on imprecise instructions. For defenders, the threat therefore combines intentional post-exfiltration data destruction with damage amplified by agent imprecision. Gambit's guidance, which CSA endorses, is that recovery planning should not end at "the database is restored." Organizations should identify the minimum set of systems needed to keep revenue flowing and verify that they can restore those systems under campaign conditions [7].

The Chainalysis case also shows a quieter form of downstream harm. The exposed fields were low sensitivity on their own. Combined with organizational role labels, however, they identify which individuals at which agencies and firms work cryptocurrency investigations. Phishing, vishing, and impersonation campaigns against those individuals are a foreseeable next step. The company's warning about impersonators offering fund recovery services suggests it shares this concern [1].

Recommendations

Immediate Actions

Organizations should begin by confirming that their detection coverage reflects the threat as it now operates. Review identity provider, cloud control-plane, and SaaS audit logs for the behaviors in the detection table above, giving priority to scripted user agents on administrative APIs, application self-assignment of privileged roles, and bulk secret reads. Where Microsoft, CrowdStrike, Gambit, or

SOCRadar have published indicators, run retrospective hunts against them [5][6][7][9]. Inventory every key and service account authorized to call external LLM inference endpoints, review provider usage logs for unfamiliar sources or anomalous volume, and alert on inference calls from hosts outside that inventory. Defenders should also treat their own AI provider keys as potential attack infrastructure, because a stolen key could fund an adversary's operation.

Customers of Chainalysis should follow the company's guidance by enforcing multi-factor authentication, ideally phishing-resistant methods, for all platform users [1][2]. They should also brief named users that impersonation attempts are likely. Help desk and password-reset processes deserve particular attention, because an attacker holding names, emails, and role labels has what it needs for convincing pretexting.

Short-Term Mitigations

Within the next 30 to 90 days, security teams should close the gap between agent tempo and response tempo. This means pre-approving automated containment actions, such as session revocation, credential rotation, and temporary account suspension, for high-confidence detections that combine multiple signals. Broader actions such as host isolation should be reserved for assets where a false positive is tolerable. Coverage outside business hours should be tested against a scenario in which an intrusion begins on a Friday night and achieves privileged access in under three hours, which matches the timelines Microsoft observed [6].

Incident response runbooks should be updated to cover the distinctive features of agent-driven intrusions. Evidence collection checklists should include agent working files, such as memory files, scope files, playbooks, and generated scripts. These should be preserved before remediation, because they may record attacker intent directly. Microsoft's guidance to preserve evidence before deleting compromised applications applies here as well [6]. Backup architecture should be reviewed to confirm that at least one copy of critical data is immutable and logically separated from production credentials, given that agents have destroyed backup tables reachable from the compromised environment [7].

Strategic Considerations

Over the longer term, organizations should revisit threat models that implicitly ranked their exposure by attacker interest. When model costs for a completed reconnaissance and exploitation scan average tens of dollars, and total model spend for a campaign that compromised at least 27 companies was estimated at $12,000 to $18,000 [7], the population of organizations worth attacking is likely to expand considerably. Mid-market firms, suppliers, and service providers with privileged connections to larger customers, as in the ConnectWise case [6], should assume that they are now within reach of adversaries who previously would have ignored them.

Patch velocity and exposure management also need to be treated as time-critical operational controls rather than periodic maintenance. Because the documented agents enumerate targets and attempt exploitation continuously and in parallel [6][7], the window between public exposure of a weakness and attempts to exploit it is likely to keep narrowing. Organizations that deploy agentic AI internally should apply the same discipline to their own agents, including least-privilege tool access, logging of every tool call, and human approval for high-impact actions. Doing so should reduce the risk that those agents are hijacked, and it helps build the telemetry practices needed to recognize hostile agents. Finally, CSA encourages organizations that experience agentic intrusions to disclose technical details, including entry vectors, dwell time, and observed agent behaviors, to the extent legally possible. The Chainalysis notice is a meaningful step toward naming the technique, but defenders across the industry would benefit from the operational specifics that remain undisclosed.

CSA Resource Alignment

This note builds on a series of CSA AI Safety Initiative publications that have tracked agentic intrusion through 2026. The white paper *Autonomous Agentic AI Adversaries* [16] established the broader argument that autonomous agentic adversaries were becoming an operational reality, and the campaign data summarized here bears out that assessment. The most directly related research note is *Agentic Intrusion Enters the Breach Record: Lessons from Chainalysis* [10], published one day before this note. That paper analyzes the Chainalysis disclosure itself and its significance as one of the few breach notices in which the victim named agentic AI as the method of intrusion. The present note complements it by placing the incident within the broader campaign data and concentrating on detection and response operations.

The detection guidance here draws on and extends *Fingerprinting Agent-Operated Intrusions: Detection and Response Guidance* [11], which catalogs behavioral fingerprints of agent operation, including tempo, hallucination artifacts, and inference-endpoint calls from compromised hosts, and recommends pre-approved automated containment. The new Microsoft and CrowdStrike reporting provides additional fingerprints, such as scripted user agents on control-plane APIs and round-the-clock operation, that fit directly into that framework. Similarly, *Agentic Intrusion Goes Retail: Detection and Hunting Lessons* [12] provides hunting guidance for the Gambit and SOCRadar campaigns summarized above, and its emphasis on recovery under campaign conditions informs the response recommendations in this note. Readers seeking more depth on the ARTEX framework behind Storm-0940 can consult *Multi-Agent Offensive Frameworks in the Wild: CyberStrikeAI and ARTEX* [17], while *LLM Agents as Post-Exploitation Tools* [18] documents an earlier in-the-wild agent-driven intrusion whose compressed timeline anticipates the response-window concerns discussed above.

For control mapping, the CSA AI Controls Matrix (AICM) v1.1 [13] provides the governing control set. Its Logging and Monitoring, Security Incident Management, and Identity and Access Management domains align with the detection, containment, and authentication recommendations above, while its Threat and Vulnerability Management domain addresses the patch velocity concerns. Organizations that deploy their own agents can use CSA's MAESTRO agentic AI threat modeling framework [14] to assess how internal agents might be hijacked or impersonated. They can then validate those controls using the techniques in CSA's *Agentic AI Red Teaming Guide* [15].

References

[1] Ax Sharma. "Chainalysis discloses data breach from 'offensive agentic techniques'." BleepingComputer, October 31, 2026.

[2] Ionut Arghire. "Chainalysis Discloses Data Breach Following AI-Driven Cyberattack." SecurityWeek, November 2, 2026.

[3] Michael Millard. "Blockchain investigator Chainalysis hit by cybersecurity incident." Cointelegraph, October 31, 2026.

[4] Eric Geller. "Agentic AI helps hackers level up, even without detailed instructions." Cybersecurity Dive, October 22, 2026.

[5] CrowdStrike Counter Adversary Operations. "CrowdStrike Uncovers New Agentic AI Campaigns." CrowdStrike Blog, October 21, 2026.

[6] Microsoft Threat Intelligence. "AI-driven, large-scale nation-state campaign using ARTEX enables supply chain compromise." Microsoft Security Blog, October 21, 2026.

[7] Eyal Sela. "AI Agents Are Hacking Online Retailers for $25 a Company." Gambit Security, September 22, 2026.

[8] Anthropic. "Disrupting the first reported AI-orchestrated cyber espionage campaign." Anthropic, November 13, 2025.

[9] Ravie Lakshmanan. "CyberXero Uses Claude Code in Campaign That Stole Over 628,000 Ukrainian Records." The Hacker News, October 12, 2026.

[10] Cloud Security Alliance AI Safety Initiative. "Agentic Intrusion Enters the Breach Record: Lessons from Chainalysis." CSA Labs, November 3, 2026.

[11] Cloud Security Alliance AI Safety Initiative. "Fingerprinting Agent-Operated Intrusions: Detection and Response Guidance." CSA Labs, October 22, 2026.

[12] Cloud Security Alliance AI Safety Initiative. "Agentic Intrusion Goes Retail: Detection and Hunting Lessons." CSA Labs, October 27, 2026.

[13] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, June 22, 2026.

[14] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 6, 2025.

[15] Cloud Security Alliance. "Agentic AI Red Teaming Guide." Cloud Security Alliance, 2025.

[16] Cloud Security Alliance AI Safety Initiative. "Autonomous Agentic AI Adversaries." CSA Labs, June 24, 2026.

[17] Cloud Security Alliance AI Safety Initiative. "Multi-Agent Offensive Frameworks in the Wild: CyberStrikeAI and ARTEX." CSA Labs, October 28, 2026.

[18] Cloud Security Alliance AI Safety Initiative. "LLM Agents as Post-Exploitation Tools." CSA Labs, June 6, 2026.