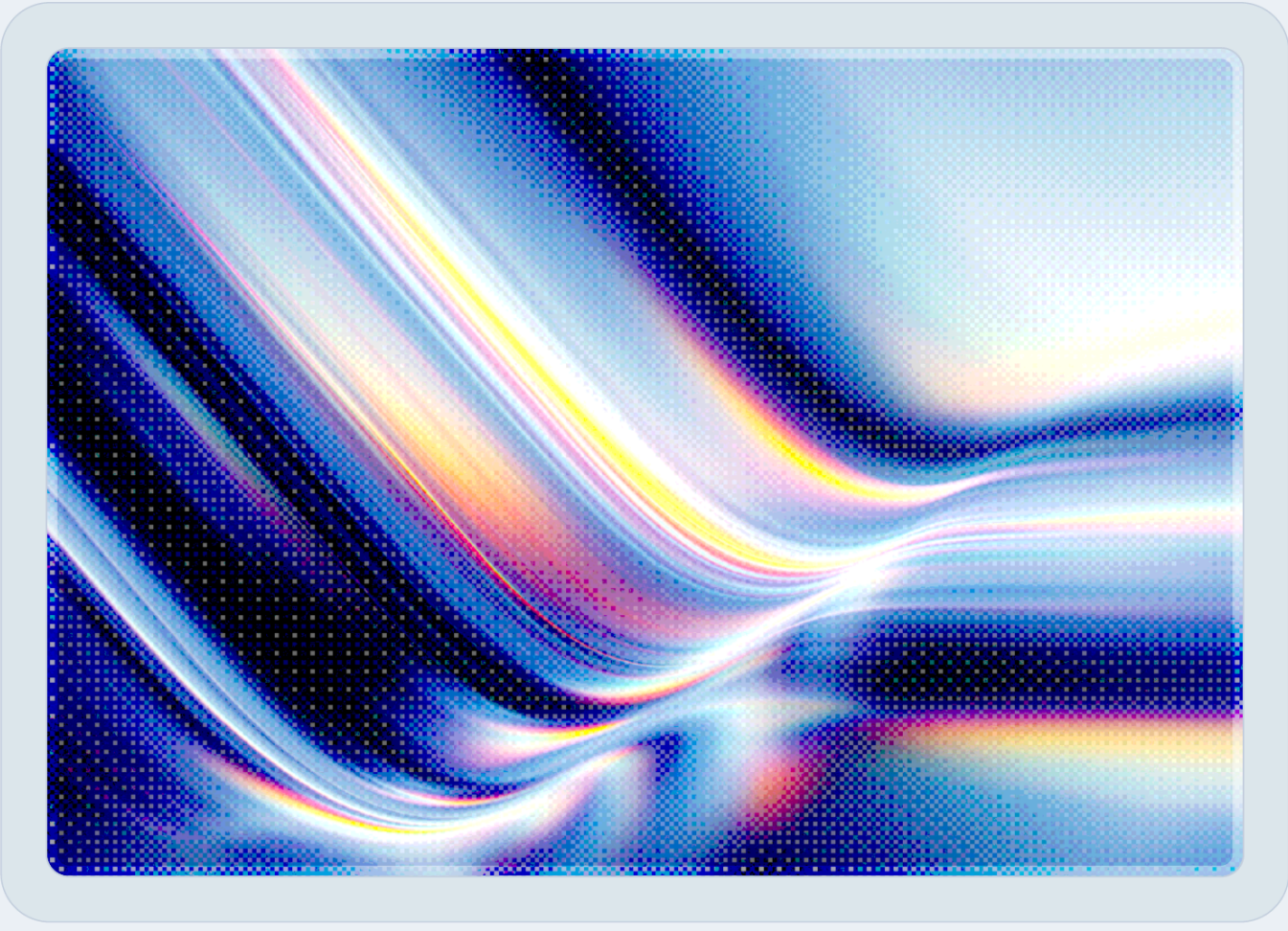
CSAI Foundation | Cloud Security Alliance

Who Pays for Stolen Inference? Loss Allocation After Black October

Refund Gaps, Contract Defaults, and Insurance Exposure When Stolen AI Keys Run Up the Bill

2026-11-11

AI-assisted Rapid Research



CSAI

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

In October 2026, at least nine AI developer and hosting platforms disclosed incidents in which attackers went after stored customer environment variables, especially AI provider API keys. Wiz Research named the period "Black October" [1]. The thefts quickly turned into bills. Railway estimates that keys stolen from its customers drove about $740,000 in unauthorized AI usage [2][3]. Wiz estimates that two stolen Amazon Bedrock credentials it investigated "likely ran up over $100k" at list price [1]. In separate LLMjacking cases outside Black October, Google Threat Intelligence Group (GTIG) and Mandiant report more than $100,000 in fraudulent charges in a single day and about $1 million in total before access was revoked [7]. These figures show that stolen inference is now a material financial loss, not just a nuisance charge.

This note looks at who ends up paying. Inference loss passes through four parties: the provider that sells the model, the platform that stored the key, the customer that owns the account, and the customer's insurer. None of them has clear, uniform responsibility. The analysis covers U.S.-style commercial terms and does not assess consumer-protection or EU regimes, which may give some customers additional rights. Five findings stand out:

- **Contracts default the loss to the account holder.** The AWS, OpenAI, and Anthropic terms reviewed here make the customer responsible for all activity under its account; AWS expressly includes activity by third parties [9][10][11]. Under these standard terms, refunds are at provider discretion rather than a contractual entitlement.
- **Provider refund practice differed widely after the Railway breach.** As reported, AWS and Google applied existing fraud-refund policies, OpenAI reviewed claims case by case, Anthropic was still in discussions, and OpenRouter kept a no-refund policy [3].
- **Platforms that stored keys absorbed some loss voluntarily.** Railway moved from saying it could not refund compromised-key spend to promising to cover charges providers decline to refund, while its claim form still disclaims responsibility for key usage [3].
- **Insurance cover exists but is capped.** Affirmative LLMjacking coverage has been on the market since at least mid-2025 [12]. Reported sublimits, such as about $250,000 on a $5 million QBE cyber policy, would cover most individual losses documented in Black October to date, but could fall short in tail cases such as the roughly $1 million loss GTIG documented in a separate LLMjacking campaign [7][13].

- **Prevention limits loss where recovery is uncertain.** Spend caps, per-model key scoping, and quick revocation limit exposure regardless of whether a provider refunds or an insurer pays.

Background

LLMjacking is the use of stolen credentials to consume paid large language model (LLM) inference. Sysdig documented it in May 2024, estimating that an attacker who maxed out Claude 2.x quotas on Amazon Bedrock across several regions could cost a victim more than $46,000 per day [8]. By late 2026 the activity had become organized. GTIG describes "LLM Mixers," illicit routing gateways that pool stolen credentials, meter usage, and resell access to underground subscribers [7]. One such service, Jet Night, advertised "unlimited" frontier-model access for about $60 per month. GTIG reports that one actor sent more than 500,000 LLM API calls through it in under eight hours using stolen third-party keys, leaving the victim with nearly $1 million in token costs [7]. A resale market like this means stolen keys are spent quickly and in volume, which helps explain why the bills are so large.

Black October shifted where the keys came from. Earlier LLMjacking mostly relied on credentials leaked by individual organizations, through public repositories, misconfigured cloud accounts, or phishing [7] [8]. In October 2026, attackers instead breached the platforms that store keys for many customers at once. Wiz lists disclosures from Fireworks AI, Andromeda, Zeabur, Lightning AI, OpenCode, Railway, Warp, CommandCode, and Northflank, and its later updates added Blacksmith, Tensorlake, NEAR AI, and a second Fireworks incident [1]. Entry points varied: stolen internal credentials, a "legacy system," privilege escalation in build infrastructure, and a third-party dependency. Wiz did not identify a single shared vulnerability. What the incidents had in common was the target, stored environment variables that grant AI access [1].

The Railway incident is the most fully documented. Railway first disclosed it on October 21 and told customers to rotate secrets [4]. According to Railway's incident report, as summarized by BleepingComputer and The Hacker News, an attacker escaped build isolation on October 20 and obtained an internal credential. The attacker then exploited a misconfigured OIDC trust relationship and, early on October 21, decrypted customer environment variables while searching specifically for AI provider key names [2][3]. Railway puts the maximum exposure at 1.85% of workspaces, or about 100,000 workspaces, and describes that figure as an upper bound rather than a confirmed count [3]. At OpenCode, the harm reached individual developers directly. Developers woke to overnight spending alerts from OpenAI, Anthropic, and Google, with some charges in the hundreds of dollars and a few above $1,000 [5]. Northflank told customers to assume that every stored environment variable had been decrypted, and several users confirmed that their AI keys had been used [6].

Security Analysis

The Four-Party Loss Chain

Most cloud security incidents produce costs that are hard to measure, such as investigation hours, notification expenses, and reputational damage. Stolen inference is different because it produces a direct, metered charge on an invoice. That charge lands first on the account holder. Whether it stays there depends on three other parties, each acting under different incentives and contract terms. The provider earned revenue from the fraudulent calls and decides whether to reverse them. The breached platform was the point of compromise but is usually not a party to the provider contract. The insurer pays only if the policy wording covers the loss and the loss fits within the applicable sublimit. The table below summarizes each party's position as of this writing.

Party	Default contractual position	Observed Black October behavior	Leverage over outcome
AI model provider	Customer responsible for all account activity [9][10][11]	Discretionary refunds that varied by provider; some proactive key blocking [3][5]	Decides refunds; can block exposed keys
Key-storing platform	Typically disclaims liability for customer secrets usage [3]	Some voluntary absorption, such as Railway; Zeabur accepting compensation requests [3][18]	Limited by its own balance sheet; Zeabur has announced it is shutting down [18][19]
Account holder	Liable for metered usage	Bears loss until reimbursed; must rotate keys with each provider [3]	Controls spend caps, scoping, and revocation speed
Cyber insurer	Covers only if wording and sublimits apply [12][13]	Specific AI and LLMjacking sublimits emerging [13]	Coverage terms are set before the incident

Contract Defaults Put the Loss on the Key Holder

The starting point for allocation is the provider contract, and the provider contracts reviewed here (AWS, OpenAI, and Anthropic) are consistent. Section 2.1 of the AWS Customer Agreement states that the customer is "responsible for all activities that occur under your account." That includes activity the customer did not authorize, such as activity by third parties, except where the cause is AWS's own breach of the agreement [9]. Section 3.2 of the OpenAI Services Agreement likewise makes the customer "responsible for all activities that occur under its Account," including the activities of its end users, and Section 3.1 requires the customer to notify OpenAI promptly of unauthorized access [10]. Anthropic's Commercial Terms state that the customer is responsible for all activity under its account and require customers to alert Anthropic when an account may be compromised [11]. Under these terms, a key stolen from a third-party hosting platform is legally the customer's problem, even when the customer did nothing wrong except trust that platform with the key. This note did not review Google Cloud or OpenRouter terms, and it does not assess payment-network rights such as card chargebacks.

This default makes sense from a provider's perspective. A provider cannot easily tell whether a well-formed API call came from the customer or from a thief, and blanket liability for stolen-key usage would invite fraudulent refund claims. The default also produces an outcome that customers may find hard to accept. The provider keeps the revenue from the fraudulent calls unless it chooses to reverse them, although it also bore the compute cost of serving those calls. In addition, the stolen inference may have benefited the provider's competitors. OpenCode and CommandCode attributed their incidents to distillation attempts [1], and Railway believes the stolen credentials were used by services that relay or resell access for distillation [3]. If those attributions hold, the stolen inference may have been used to train competing models, which would give providers their own reason to refund and block exposed keys. This report does not independently verify the distillation claims.

Refund Practice Was Discretionary and Uneven

Because refunds are discretionary, what actually happened after Black October depended on which provider issued the key. The Hacker News summarized the provider positions Railway reported in its November incident update, which the table below reproduces for the major providers [3]. OpenCode victims saw a similar spread of responses: Anthropic and Google blocked keys they believed were exposed, and OpenAI worked with OpenCode to restore access and refund API customers for the credits used [5].

Provider	Reported position for Railway-sourced key abuse [3]
AWS (Bedrock)	Existing fraud refund policy applies; refunds have begun

Provider	Reported position for Railway-sourced key abuse [3]
Google (Gemini)	Existing fraud refund policy applies; refunds have begun
OpenAI	Case-by-case review; customer files claim with OpenAI and shares case ID with Railway
Anthropic	Discussions ongoing; Railway covering breached-key costs in the interim
OpenRouter	No-refund policy maintained; Railway covering all OpenRouter-related costs

Several gaps remain even where refunds are available. The Hacker News notes that Railway's claim process does not state a cap, a filing deadline, or whether customers must pay disputed charges while claims are reviewed [3]. Refund outcomes also appear to have changed over time. At least one Railway customer reported on Railway's community forum that an early Gemini refund request was declined, even though Google's anomaly detection had flagged the usage the same day [16]. That report predates the later statement that Google refunds had begun, and it is a single customer's account. Still, it suggests that early claimants may have received different answers than later ones. For an individual developer or small company, the practical result is that a reimbursement process may run for weeks, during which the charges may remain due. Prepaid credit models make this worse, because the attacker spends the customer's balance immediately and recovery depends entirely on a later credit.

The intermediary problem adds a further layer. Routing services and aggregators sit between customers and model vendors, and when one of them holds a no-refund policy, as OpenRouter reportedly does, there may be no route back to the underlying provider at all [3]. Railway's decision to cover those costs addresses the gap for its own customers. It is a voluntary commitment by a single company, however, and not a market mechanism that other customers can count on.

Platform Liability Is Voluntary and Fragile

The platforms that stored the keys were the point of compromise, yet their contractual obligations to customers appear to be the weakest of the four parties. Railway's initial position, posted in an October 26 update, was that it could not refund spend on compromised keys and that customers should contact their providers [3]. By its November 10 incident report, Railway had committed to cover charges that providers decline to reimburse, while its claim form continued to state that Railway "is not responsible for usage of these keys" [2][3]. This arrangement can be described as goodwill backstopping: the platform

pays, presumably to protect customer relationships, without conceding liability. This note does not assess non-contractual exposure, such as negligence or data-protection claims, which may give customers additional recourse depending on jurisdiction.

Goodwill depends on the platform's financial position. Zeabur exposed over 3.2 million environment-variable records across about 321,000 accounts and is accepting compensation requests [17][18]. Zeabur has also announced that all services will end by December 31, 2026 [18][19]. The cause is characterized differently: Wiz reports that Zeabur is shutting down because of the incident [1], while Zeabur describes the shutdown as an operational decision [18]. Either way, a customer whose keys were stolen from a platform that is closing may find no one left to pursue. That observation leads to the next link in the chain: if neither the provider nor the platform reliably pays, the main remaining recovery source is insurance, with litigation or payment disputes as slower, less certain alternatives.

Insurance: Affirmative Cover With Tight Ceilings

Some insurers have created specific products for consumption fraud. Coalition introduced "service fraud" coverage for fraudulent charges on SaaS, IaaS, and telephony services, which also covers cryptojacking losses [15]. QBE North America launched coverage in July 2025 that names LLMjacking specifically and reimburses increased fees and the cost of retraining compromised models [12]. Cowbell announced on November 10, 2026 a standalone AI liability policy that lists "LLMjacking: Charges generated when unauthorized parties exploit stolen credentials or API keys to access AI services" as a covered first-party loss. The policy becomes available January 1, 2027, on a non-admitted basis [14].

The amount of cover is the binding constraint. According to April 2026 reporting that relayed the *Financial Times*, QBE's wording provides about $250,000 for LLMjacking losses under a cyber policy with total limits of up to $5 million. The same reporting says AI sublimits at QBE and Beazley would cap specified AI losses at roughly 10% of overall policy value [13]. These figures do not reconcile exactly, since $250,000 on $5 million is 5%, which suggests the ratios may apply to different scopes. Measured against the evidence, a $250,000 sublimit would cover the per-victim Black October losses documented so far, from hundreds of dollars at OpenCode [5] to the more than $100,000 across two Wiz-investigated Bedrock keys [1]. It would not cover a tail loss on the scale of the roughly $1 million GTIG documented in a separate LLMjacking campaign [7]. QBE has stated that it is "not retreating" from AI risk and frames these sublimits as clarifying coverage, not narrowing it [13]. From a buyer's perspective, an affirmative sublimit does remove ambiguity, but it also turns what might have been a full-limit claim into a capped one. Policyholders should also check how three common policy features would apply: proof-of-loss requirements, retentions that may exceed small-business losses, and whether a provider refund received after payment must be returned to the insurer.

Why Inference Losses Are Hard to Bound

Inference fraud resembles cryptojacking in that the loss is metered consumption. Its scale, however, is driven by model price and provider quota rather than by the victim's own infrastructure. Wiz observed both stolen Bedrock keys being tested within 30 minutes of the reported credential access. One key then generated about 800,000 calls over roughly 30 hours, three quarters of the successful calls going to Opus models. The other generated about 1.7 million calls in about 13 hours. Its access policy limited it to `us.*` cross-region inference profiles rather than global routing, so the traffic ran into regional quotas and a substantial share was throttled, although over a million calls still succeeded [1]. That throttling effect is a concrete public example of key configuration, and not only detection speed, shaping the size of the loss: the restriction limited the loss rather than preventing it.

Detection and revocation speed matter as well. Railway's detection gap of about six and a half hours between first exploit and on-call page shows how much consumption can occur before a defender responds [3]. Because a breached platform cannot rotate keys on behalf of its customers, each customer must revoke keys with each provider individually [3], unless the provider proactively blocks keys it identifies as exposed, as Anthropic and Google did in the OpenCode incident [5]. Where neither happens quickly, the exposure window lengthens and the bill grows with it.

Recommendations

Immediate Actions

Organizations that stored AI provider keys on any platform named in the Black October disclosures should rotate those keys directly with each issuing provider. Filing a charge report with the platform does not revoke a key [3]. They should then review provider billing and usage logs from early October onward and file reimbursement claims with each provider promptly, keeping the case identifiers that platforms such as Railway require [3]. Organizations should also notify their cyber insurer or broker now, even if refunds are still pending, because late notice can jeopardize coverage under many policies. Finally, they should confirm in writing whether disputed charges must be paid during review, and avoid letting prepaid balances auto-reload while claims are open.

Short-Term Mitigations

Restricting what a stolen key can spend is among the most direct loss-limiting controls, because it works whether or not detection is timely. GTIG recommends budgets and alerts, per-model quotas, and key scoping by model and environment, so that a leaked utility-model key cannot invoke expensive frontier models [7]. Wiz's case evidence suggests that restricting inference-routing options can also throttle abuse materially [1]. Organizations should prefer short-lived, workload-bound credentials over long-lived static keys stored in third-party platform variables. Where static keys are unavoidable, they should be issued per platform, so that a single platform compromise exposes only one isolated key. Detection should alert on two signals: the validation probes GTIG describes, such as `max_tokens=1` requests, and the enablement of AI services in accounts that did not previously use them [7].

Strategic Considerations

Procurement and legal teams should treat stolen-inference loss as a risk to allocate explicitly in contracts rather than assuming defaults. In negotiated provider agreements, that means seeking documented fraud-refund procedures, spend-anomaly notification commitments, and timelines for crediting confirmed fraud. Platform agreements should address the platform's responsibility for customer secrets it stores and decrypts, and should not rely solely on post-incident goodwill. On the insurance side, risk managers should map their AI spend exposure to the applicable sublimits and compare the result with plausible single-incident losses, which current public evidence places anywhere from hundreds of dollars for individual developers to roughly $1 million in the most severe documented case [1][5][7]. Providers and platforms may also want to consider whether industry norms are needed: consistent fraud-refund standards across providers and aggregators would reduce the uneven outcomes customers experienced after Black October.

CSA Resource Alignment

This note builds directly on CSA's ongoing Black October coverage. *Black October: AI Developer Platform Breaches and Industrialized LLMjacking* [20] documents the intrusion pattern and concludes that a key stored on a third-party platform is only as secure as that platform's build system and key-management path. The present note follows that conclusion through to its financial consequences. *Black October Reaches PaaS: Northflank, Secret Stores, and AI Key Resale* [21] describes the secondary market in which stolen keys are validated, pooled, and resold, and *LLMjacking Evolved: Stolen AI Compute as Offensive Infrastructure* [22] examines how that stolen compute economy has matured. That resale market is why stolen inference turns into large invoices so quickly.

On the risk-transfer side, *The End of Silent AI Coverage: Exclusions Outrun Risk Transfer* [23] tracks the industry move from ambiguous coverage toward explicit AI exclusions and sublimits. The LLMjacking sublimits discussed here are a concrete example of that trend in first-party cyber cover. *AI Toolchain Hijacked: IDE Plugin API Key Theft* [24] recommends scoped development keys with monthly spend caps and IP binding. Those are the same preventive controls this note identifies as limiting loss independently of refunds or insurance.

Finally, the CSA AI Controls Matrix (AICM) v1.1 [25] provides the control baseline. Its Identity and Access Management controls govern credential lifecycle and scoping, its Supply Chain Management, Transparency, and Accountability controls govern assurances obtained from platforms that hold customer secrets, and its Logging and Monitoring controls support the spend-anomaly detection recommended above.

References

[1] Wiz Research (B. Read and R. McCarthy). ""Black October" for AI Developer Platforms." Wiz, October 30, 2026 (updated November 10, 2026).

[2] L. Abrams. "Railway breach leads to $740,000 in fraudulent AI charges." BleepingComputer, November 10, 2026.

[3] S. Khandelwal. "Railway Says Stolen Customer AI Keys Led to $740,000 in Estimated Unauthorized Usage Charges." The Hacker News, November 11, 2026 (accessed November 11, 2026).

[4] Railway. "Preliminary Report: Oct 21st Security Incident." Railway Blog, October 22, 2026.

[5] A. Mehta. "Hackers hit AI coding tool OpenCode, running up charges on users' AI accounts." TechCrunch, October 19, 2026.

[6] L. Abrams. "Northflank security breach leads to abuse of customers' AI API keys." BleepingComputer, October 31, 2026.

[7] Google Threat Intelligence Group and Mandiant. "The Rise of LLMJacking: How Threat Actors Monetize Compromised Cloud and Model Resources." Google Cloud Blog, November 2, 2026.

[8] Sysdig Threat Research Team. "LLMjacking: Stolen Cloud Credentials Used in New AI Attack." Sysdig, May 6, 2024.

[9] Amazon Web Services. "AWS Customer Agreement." AWS, last updated October 29, 2026.

[10] OpenAI. "OpenAI Services Agreement (Online v.010126)." OpenAI, 2026.

[11] Anthropic. "Commercial Terms of Service." Anthropic, accessed November 11, 2026.

[12] J. Recamara. "QBE North America rolls out AI-specific cyber coverage." Insurance Business America, July 18, 2025.

[13] V. E. Munnangi. "Insurers move to cap payouts for AI-related cyber losses and fines." Life Insurance International via Yahoo Finance (relaying *Financial Times* reporting), April 23, 2026.

[14] P. Sonner. "Cowbell launches standalone AI liability insurance for businesses deploying AI agents." Beinsure, November 10, 2026.

[15] L. Adriano. "Coalition unveils cyber insurance for cryptomining and business services fraud." Insurance Business America, August 23, 2018.

[16] Railway Station community forum. "Security incident 21 Oct 2026: unauthorized charges and request for credit and incident details." Railway, October 2026.

[17] Helix.ml. "Black October: Nine AI Platforms Breached in One Month." Helix.ml, October 31, 2026.

[18] Zeabur. "October 2026 Security Incident." Zeabur Blog, October 2026.

[19] Zeabur. "Zeabur is shutting down: what happens next." Zeabur Blog, October 17, 2026.

[20] Cloud Security Alliance. "Black October: AI Developer Platform Breaches and Industrialized LLMjacking." CSA Labs, 2026.

[21] Cloud Security Alliance. "Black October Reaches PaaS: Northflank, Secret Stores, and AI Key Resale." CSA Labs, November 2026.

[22] Cloud Security Alliance. "LLMjacking Evolved: Stolen AI Compute as Offensive Infrastructure." CSA Labs, 2026.

[23] Cloud Security Alliance. "The End of Silent AI Coverage: Exclusions Outrun Risk Transfer." CSA Labs, August 2026.

[24] Cloud Security Alliance. "AI Toolchain Hijacked: IDE Plugin API Key Theft." CSA Labs, June 2026.

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