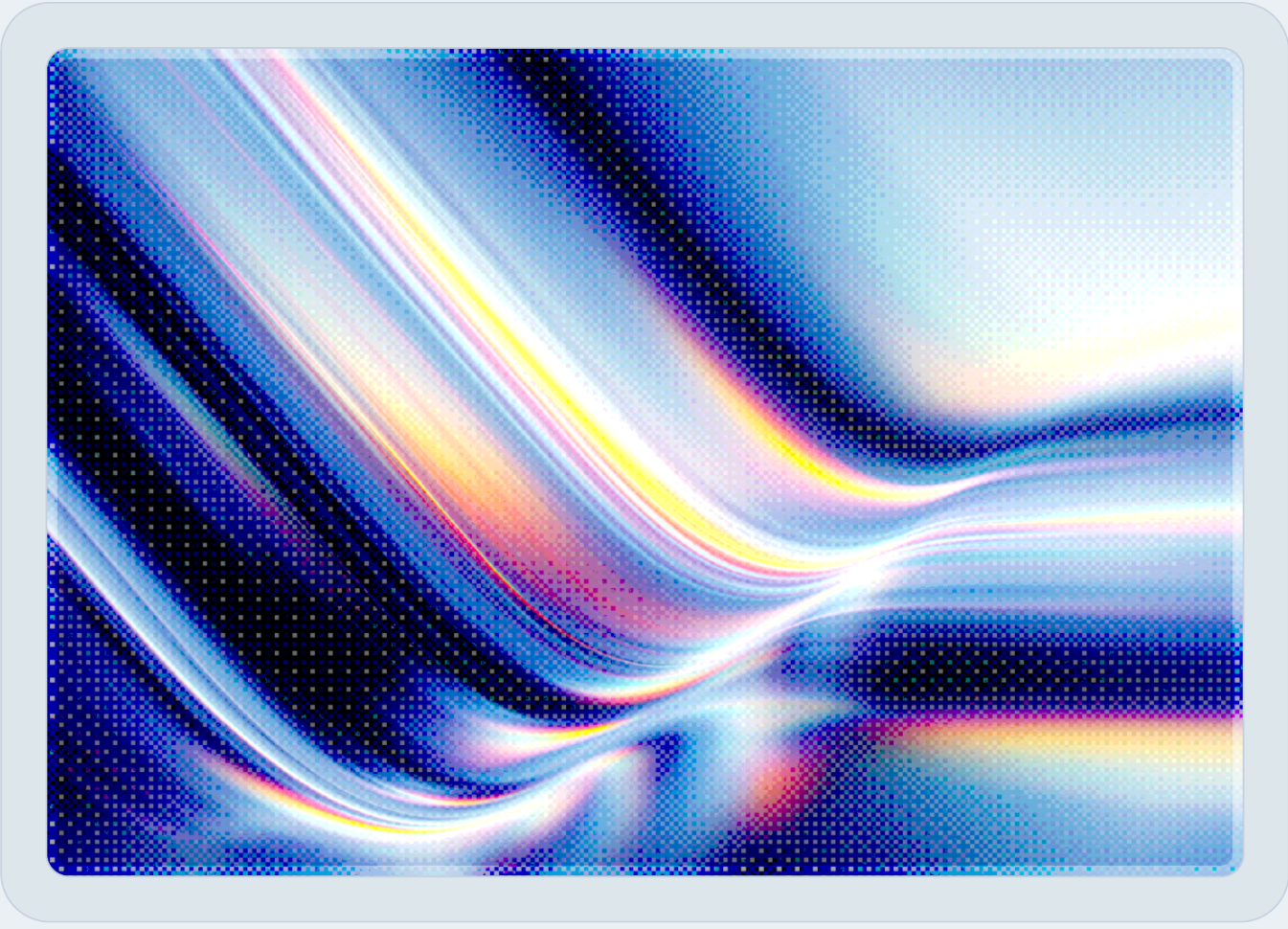
CSAI Foundation | Cloud Security Alliance

Fireworks AI Breach: Inference Providers as Supply-Chain Chokepoints

Security Implications and Guidance for the Downstream Cascade

2026-11-17

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

Fireworks AI, a hosted inference and fine-tuning platform for open-weight models, has disclosed two security incidents in five weeks. It published the second as an update to the first, but public reporting has not established whether the two share an actor [2][3]. The first incident, detected on October 6, 2026, involved misuse of internal credentials to read environment variables, including customer API keys and tokens stored as secrets, from evaluation jobs in a limited number of accounts [1]. The second began on October 23, when a third-party cloud provider hosting Fireworks research and development machines was compromised. The attacker then reached Fireworks development machines on October 24. According to Fireworks' November 9 update, the attacker accessed account and billing information and, for a limited set of accounts, secrets, datasets, and incomplete model weights [2][3].

The second incident reached people who had never dealt with Fireworks directly. On November 11, Notion disclosed that Notion AI chat transcripts used for its Insights feature were in the exposed data, affecting fewer than 1% of Notion workspaces [4][5]. Notion said Fireworks had contractually committed to zero data retention (ZDR), yet "data was retained from the Fireworks endpoints that we used" [4]. That gap between a contractual retention promise and the data that actually sat at rest is a central lesson of this incident.

This note draws four conclusions for security leaders. First, inference providers are both *credential reservoirs* and *data processors*, so a breach can expose the secrets customers store on the platform as well as the content customers send through it. Second, ZDR should be verified feature by feature rather than accepted at the vendor level. Fireworks' own data-retention policy acknowledges that stored responses, batch jobs, and training retain content even though its core inference product is ZDR by default, although neither company has disclosed which of these features held Notion's transcripts [6]. Third, the blast radius extends through fourth parties: end users of a SaaS product can have their data exposed by their provider's AI vendor, a company they never chose, never contracted with, and likely did not know existed. Finally, credential exploitation moves much faster than vendor notification. In the Black October cases Wiz investigated, which have not been shown to involve Fireworks, stolen keys were first used within 30 minutes of credential access [7].

Background

Fireworks AI runs hosted inference, fine-tuning, and evaluation for open-source models such as Llama, Mistral, and Gemma [3]. Customers use it to serve models through an API and to run batch and training jobs. Like other platforms in this category, it occupies a dense position in the AI supply chain: it holds credentials its customers give it so their jobs can reach other services, it processes the prompts and datasets those customers send, and it stores fine-tuned model artifacts on their behalf. These roles concentrate value, which makes such platforms attractive targets.

The first incident fits a wider pattern Wiz Research named "Black October." Wiz counted at least nine AI developer platforms that disclosed security incidents in October 2026, Fireworks among them, and later added more names to that list [7]. Attackers repeatedly went after environment variables, especially AI provider keys. Fireworks' initial bulletin said an unauthorized third party used internal Fireworks credentials to read evaluation-job environment variables and that API keys and tokens stored as secrets were likely captured in certain accounts [1]. Fireworks revoked internal credentials, affected API keys, and keys the attacker had created. Its bulletin also warned customers that "removing a stored secret alone does not invalidate the underlying credential" [1].

The second incident came through a different path. A third-party cloud provider hosting Fireworks R&D machines had a security incident on October 23. The next day, the unauthorized party used that access to reach Fireworks development machines and ran an unauthorized workload that "caused additional customer data impact"; Fireworks says it detected this activity the same day [2]. Fireworks has not named the cloud provider. It brought in Palo Alto Networks Unit 42 for forensics, disabled self-serve dataset uploads and downloads for Batch Inference and dataset-based training during the investigation, and said it detected no impact to its Inference, Training SDK, or Nexus products, which it describes as ZDR by default [2][3]. Fireworks' update links the activity to a financially motivated group that sells unauthorized LLM access to illicit routing gateways, a practice known as LLMjacking [2]. Public reporting does not settle whether both incidents were the work of the same actor.

The downstream disclosure followed. According to Notion, Fireworks alerted it on November 2 to an exposed dataset that Notion first believed held only synthetic test data. Notion disconnected from Fireworks by November 3, and Fireworks confirmed on November 8 that the dataset included Notion AI chat transcripts used for the Insights feature [4][5]. Notion said most affected transcripts were "unattributable," meaning they carried no metadata linking them to a user or workspace. It also warned that the chat text itself may contain names, email addresses, API tokens, or text selected on the current page [4]. Notion disabled Insights, stopped serving open-weight models through Fireworks, and announced an audit of vendor ZDR and data minimization practices [4].

Security Analysis

The Inference Provider as a Dual Chokepoint

Earlier CSA analysis of Black October treated AI build-and-host platforms mainly as *credential reservoirs*: places where many customers' secrets accumulate, so that one breach exposes all of them at once [11]. The first Fireworks incident fits that model. The second one adds a further dimension. An inference and fine-tuning provider is also a *data processor* that handles prompts, evaluation sets, and training corpora, and that stores derived artifacts such as fine-tuned weights. An attacker who gets in can take credentials that enable further intrusion, content that may itself contain secrets or personal data, and intellectual property in the form of customer datasets and partially trained models.

These exposures have different consequences and require different responses. The table below sets out the categories Fireworks reported and what each means for a downstream organization.

Exposure category	What Fireworks reported	Downstream consequence	Primary response
Stored secrets (environment variables)	First incident: API keys and tokens in evaluation-job environment variables [1]. Second incident: secrets for a limited set of accounts [2]	Lateral access to other providers; LLMjacking billing fraud	Rotate at the issuer, not just on the platform
Datasets	Datasets for certain customers [2]	Exposure of personal data, business content, or third-party user data	Data-breach assessment; notify downstream users
Incomplete model weights	Subset of incomplete weights for certain customers [2]	Loss of fine-tuning IP; possible leakage of memorized training data	IP risk review; assess memorization exposure
Billing and account metadata	Addresses, invoice history, balances, resource names [2]	Targeted phishing; reconnaissance of AI projects	User awareness; review resource naming

Zero Data Retention Is a Feature Setting, Not a Vendor Attribute

The most important finding for enterprise buyers is the gap between Notion's understanding of its contract and what was actually stored. Notion says Fireworks had contractually committed to ZDR for Notion data, yet data was retained from the endpoints Notion used [4][5]. Fireworks, for its part, says its inference product is ZDR by default, and its own published data-retention policy, as quoted by VentureBeat, states that "a few features do store customer content, such as stored responses, batch jobs, and training" [6]. Notion used Fireworks for classification jobs supporting Insights, and its notice also cites experiments training task-specific models with data from volunteers in an early-access program [4]. Neither company has publicly said which Fireworks feature held the transcripts [6].

The precise mechanism remains unconfirmed, and other explanations, such as data copied into research environments or a misconfiguration, have not been ruled out. Even so, the incident shows a structural weakness in how ZDR is often bought and verified. A buyer negotiates a retention commitment at the vendor level, and engineering teams then adopt whichever endpoints suit the job, including batch, evaluation, and training workflows that keep data at rest by design. VentureBeat's reporting suggests that on Fireworks a ZDR policy setting may be off unless explicitly configured and that enabling it does not delete content already stored [6]. If those observations hold more broadly, procurement-level assurances can drift quietly out of line with how the product is actually used. We infer that similar gaps probably exist at other inference providers that offer both real-time and batch or training services, though this incident alone does not show how common they are.

The Fourth-Party Cascade

The Fireworks incident affected parties at four levels, and each level had progressively less visibility into, or control over, the original failure. The chain below runs from the unnamed cloud provider to individual Notion users.

Tier	Party	Relationship to the failure	Visibility and control
0	Unnamed third-party cloud provider	Initial compromise, October 23 [2]	Not publicly disclosed
1	Fireworks AI	Development machines reached via the provider, October 24 [2]	Direct; notified customers

Tier	Party	Relationship to the failure	Visibility and control
2	Fireworks customers (e.g., Notion)	Data and secrets held by Fireworks exposed [2][4]	Contractual; dependent on vendor notice
3	End users and workspaces of those customers	Chat transcripts sent to Fireworks by Notion [4]	Minimal; dependent on Notion's subprocessor disclosures; likely unaware of Fireworks

The table shows that the people whose data was exposed had no say in the vendor decision that put it at risk. Notion users chose Notion, not Fireworks, and the transcripts reached Fireworks through an analytics feature (Insights) that processed users' chats, not through any choice by those users to use Fireworks [5]. This dynamic resembles OAuth-token SaaS supply-chain compromises such as the 2025 Salesloft/Drift breach, where trust granted to an integrated vendor cascaded across hundreds of downstream customer environments [15]. It is sharper in the AI case, though, because what flows to inference vendors is often unstructured conversational text that users treat as private and that can include credentials pasted mid-conversation. Notion's advice that customers rotate any credentials typed into chats acknowledges this directly [4].

Timing compounded the cascade. Fireworks detected the intrusion into its development machines on October 24 [2]. It first alerted Notion to an exposed dataset on November 2, nine days later, and confirmed on November 8 that the dataset contained Notion AI chat transcripts [4][5]. Intervals of this length may be normal for complex forensics, and Fireworks did alert Notion before its investigation had confirmed the scope. The intervals still matter because credential abuse moves much faster. In other Black October cases, not shown to involve Fireworks, Wiz observed stolen keys being validated within 30 minutes and then used for about 2.5 million unauthorized Amazon Bedrock calls across two cases, likely costing more than $100,000 at list price [7]. Google Threat Intelligence Group and Mandiant separately reported LLMjacking charges above $100,000 in a single day at some victims and cumulative charges reaching $1 million before access was revoked [8]. When credentials are part of the exposure, the downstream organization cannot wait for the vendor's investigation to finish.

Concentration and the Shadow Vendor Problem

Fireworks is one of many providers serving open-weight models, but its position in the stack gives it outsized reach. The Hacker News reported that Notion's disclosure followed Fireworks' expanded advisory, which covered customer datasets and model weights [9]. VentureBeat, citing reporting by The Record, put Fireworks' customer count above 10,000 [6]. Inference providers commonly serve software companies that embed AI features in their own products, as Notion did; where that holds, a single

provider compromise can reach end users across many unrelated applications. Earlier CSA work has argued that concentration is a central driver of AI supply-chain risk, with shared control points creating industry-wide blast radius [11][12]. The Fireworks cascade is an example of that dynamic at the inference layer specifically.

A related problem is visibility. Many organizations route traffic to open-weight models through secondary providers to save money or reduce latency, often as an engineering decision that never reaches vendor risk review. Notion's decision to stop serving open-weight models through Fireworks shows that the dependency extended beyond the Insights feature, yet secondary inference routes of this kind are exactly the dependencies that subprocessor lists and procurement inventories tend to miss.

Recommendations

Immediate Actions

Organizations that use Fireworks AI, directly or through a SaaS product, should act on the assumption that any secret stored on the platform before late October 2026 may be compromised. That bound is deliberately conservative because the start date of the first incident's unauthorized access has not been disclosed. Credentials should be rotated at the issuing provider, because deleting a secret from Fireworks does not invalidate it [1]. Teams should review usage and audit logs at each issuing provider (cloud, model provider, source control, and data stores) from October 1, 2026 onward, the window recommended in CSA's Black October guidance [10]. Those reviews should look for the low-volume validation calls GTIG describes, such as `max_tokens=1` or zero-temperature pings, and for first-time calls from unfamiliar networks [8][10].

Organizations should also inventory every inference provider currently receiving their data, including providers reached through gateways, routers, and feature-level integrations owned by product teams. For each one, confirm which specific endpoints and features are in use and whether any of them (batch inference, evaluation, stored responses, fine-tuning, or dataset uploads) keep content at rest. Organizations that are Notion customers and were notified should follow Notion's guidance: download and review affected transcripts in a restricted location, rotate any credentials found in them, and watch for phishing that uses details from those conversations [4].

Short-Term Mitigations

Over the next one to three months, organizations should turn ZDR from a contractual statement into a verified, per-feature control. Vendor questionnaires and contracts should list the specific endpoints covered by ZDR, require the provider to state which features retain data and for how long, and require written confirmation of the account-level policy setting. Where a provider offers a ZDR configuration flag, security teams should check it and verify that existing stored content has been deleted. Enabling a policy going forward does not remove data already at rest [6].

Credential exposure on inference platforms should be reduced at the source. CSA's Black October guidance recommends issuing separate keys per platform and environment, scoping keys to the models they need, setting hard spending and rate limits instead of alert-only budgets, and replacing static keys with workload identity federation or short-lived tokens where possible [10][11]. Organizations should also set an internal revocation target measured in hours for multi-platform credential exposure and rehearse it, since vendor notification may arrive days or weeks after exploitation has begun.

Strategic Considerations

Organizations that send customer data to inference providers should treat those providers as subprocessors with the same disclosure, notification, and audit obligations that apply to other data processors. They should be listed in public subprocessor disclosures, covered by breach-notification clauses with fixed timelines, and included in incident tabletop exercises. Product teams should apply data minimization before content leaves the organization, removing secrets and personal data from prompts and analytics datasets, because, as Notion's guidance to rotate credentials typed into chats acknowledges, users paste secrets into AI conversations regardless of policy [4].

Assurance programs also need to cover the inference provider's own supply chain. In this case the initial compromise was at a cloud provider hosting R&D machines, a fourth party to Fireworks' customers and a fifth party to Notion's users. Vendor assessments should ask whether R&D and development environments are separated from production customer data, whether customer datasets are ever copied into research environments, and how quickly the provider can determine which customers' data a compromised system could reach. These questions are not specific to AI, but the volume and sensitivity of conversational data flowing to inference providers make them more pressing.

CSA Resource Alignment

CSA's **Black October: AI Developer Platform Breaches and Industrialized LLMjacking** research note (November 4, 2026) covers the first Fireworks incident alongside the wider October wave [10]. Its immediate and short-term guidance on rotating at the issuer, reviewing provider-side logs from October 1, and scoping and capping AI keys applies directly to Fireworks customers. This note extends that analysis to the second incident, which showed that the same platforms also hold customer content and model artifacts, not only secrets.

The companion note **Black October: Concentration Risk in the AI Build-and-Host Layer** (November 5, 2026) describes build and hosting platforms as credential reservoirs and recommends building a "credential reservoir map" of every external platform that holds organizational secrets [11]. The Fireworks cascade suggests that map should become a *data and credential* reservoir map, recording for each provider which features retain content and whose data (including end users') flows through them. The one-to-two-week interval between the intrusion and confirmed downstream impact is consistent with that note's finding that vendor notification can trail exploitation.

CSA's whitepaper **AI Infrastructure Monoculture: Foundation-Layer Concentration Risk** (April 2026) provides the structural framing: shared AI infrastructure components create correlated failure modes across many dependent organizations [12]. Inference providers serving open-weight models are a clear instance of such a shared layer. The research note **The OAuth Gap: AI SaaS Supply Chain Blast Radius** (May 2026) examines the parallel cascade pattern in OAuth-integrated AI SaaS tools, where one compromised integration reaches many downstream customer environments [15].

On controls, the **AI Controls Matrix (AICM) v1.1** supplies the relevant domains: Supply Chain Management, Transparency and Accountability (STA) for subprocessor oversight; Data Security and Privacy Lifecycle Management (DSP) for retention and minimization; Identity and Access Management (IAM) for credential scoping and rotation; and Security Incident Management, E-Discovery, and Cloud Forensics (SEF) for notification timelines [13]. Organizations assessing inference providers should ask for **STAR** registry entries and AI-specific assurance artifacts, and should explicitly test the per-feature ZDR claims described above rather than accepting vendor-level attestations [14].

References

[1] Fireworks AI. "October 2026 Security Incident." Fireworks AI Security Bulletins, October 10, 2026.

[2] Fireworks AI. "Update on October 2026 Security Incident." Fireworks AI Security Bulletins, November 9, 2026.

[3] Gatlan, Sergiu. "Fireworks AI Discloses New Breach After Cloud Provider Hack." BleepingComputer, November 10, 2026.

[4] Notion. "Notice About the Fireworks AI Incident That May Impact Some Notion Workspaces." Notion Help Center, November 2026.

[5] Toulas, Bill. "Notion Says Fireworks Breach Exposed Customer AI Chat Transcripts." BleepingComputer, November 11, 2026.

[6] Columbus, Louis. "Attackers Breached the Fireworks AI Storage Where Notion AI Chats Sat Outside Zero Data Retention." VentureBeat, November 11, 2026.

[7] Wiz Research. "“Black October” for AI Developer Platforms." Wiz Blog, October 30, 2026 (updated through November 10, 2026).

[8] 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.

[9] Lakshmanan, Ravie. "Fireworks AI Confirms Wider Breach Affecting Customer Datasets, Model Weights and Notion AI Chats." The Hacker News, November 10, 2026.

[10] Cloud Security Alliance AI Safety Initiative. "Black October: AI Developer Platform Breaches and Industrialized LLMjacking." CSA Labs, November 4, 2026.

[11] Cloud Security Alliance AI Safety Initiative. "Black October: Concentration Risk in the AI Build-and-Host Layer." CSA Labs, November 5, 2026.

[12] Cloud Security Alliance AI Safety Initiative. "AI Infrastructure Monoculture: Foundation-Layer Concentration Risk." CSA Labs, April 2026.

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

[14] Cloud Security Alliance. "Security, Trust, Assurance and Risk (STAR) Registry." Cloud Security Alliance.

[15] Cloud Security Alliance AI Safety Initiative. "The OAuth Gap: AI SaaS Supply Chain Blast Radius." CSA Labs, May 1, 2026.