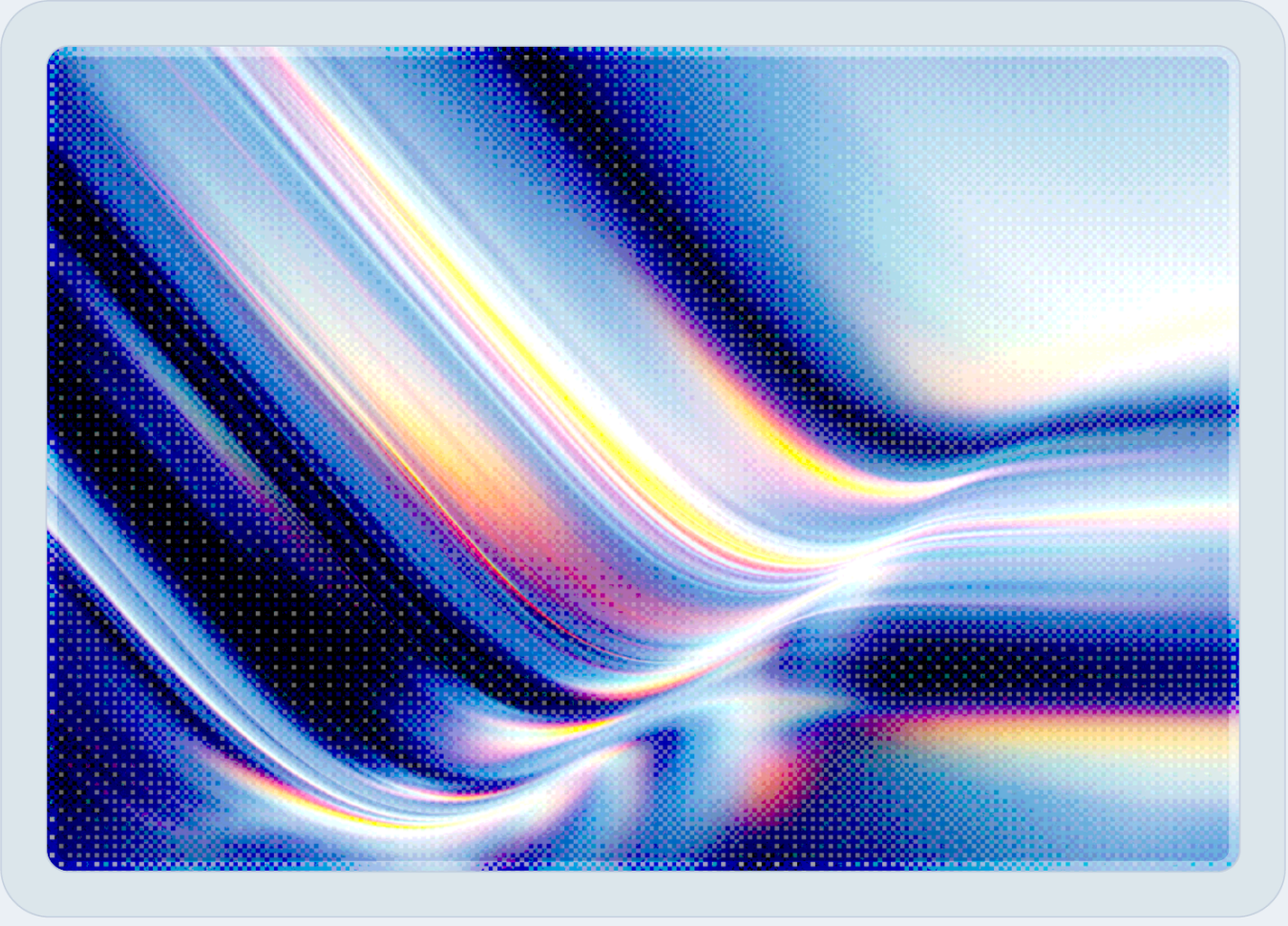
CSAI Foundation | Cloud Security Alliance

Your Vendor's Inference Provider: Fourth-Party AI Subprocessor Risk

Security Implications and Guidance After the Fireworks–Notion Exposure

2026-11-12

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 November 2026, Notion disclosed that Notion AI chat transcripts from fewer than 1% of its workspaces had been exposed. The breach took place at Fireworks AI, an inference and fine-tuning platform that Notion used to serve open-weight models and to process data for its AI usage "Insights" feature [3][4][5]. Notion said its own systems were not breached [5]. The data was exposed one layer down, at a provider that most Notion customers had probably never evaluated, contracted with, or heard of. For those customers, Fireworks was a fourth party: their vendor's vendor.

The incident matters for three reasons beyond the parties involved. First, zero data retention (ZDR) commitments were in place and data was still exposed. Notion has stated that Fireworks had contractually committed to ZDR for Notion data [3], and Fireworks describes its inference service as ZDR by default [2]. The exposed material was nonetheless held at rest in a dataset. Neither company has explained how, but the most likely explanation is that this processing path fell outside the scope of those commitments. Second, the attack reached Fireworks through yet another party, a third-party cloud provider hosting Fireworks research and development machines [2]. From the perspective of a Notion customer, the compromise therefore entered at what third-party risk management convention would call a fifth party. Third, Fireworks attributes the activity to a financially motivated group monetizing unauthorized LLM access [2]. That is the same economy Google Threat Intelligence Group and Mandiant described in their November 2026 "LLMJacking" report [7]. Inference providers hold credentials, datasets, and model weights, and this incident suggests they are becoming direct targets of that economy.

Security and privacy leaders should take four points from the incident. Vendor reviews should ask which inference providers sit behind each AI feature, not only which foundation models are used. "Zero data retention" should be treated as a scoped claim that applies to particular processing paths, and evaluation, analytics, batch, and fine-tuning workflows should be examined separately. Users debugging configuration files or API calls in an AI assistant can paste credentials into it, so organizations should plan for exposed AI chat logs to contain secrets. Finally, the visibility obligations already in data protection law and supply chain guidance should be extended explicitly to AI subprocessors [8][9][10].

Background

The incident sequence

The public record comes mainly from Fireworks' security bulletins and from Notion's statements as reported by the trade press. It shows two related intrusions at Fireworks over roughly a month. Fireworks' investigation is ongoing, and some of the facts summarized here may change as it continues. On October 6, 2026, Fireworks identified an unauthorized party using internal Fireworks credentials. That party had accessed environment variables used in customer evaluation jobs, including API keys and tokens stored as secrets for a limited number of accounts [1]. In its original bulletin, Fireworks stated that it had not detected unauthorized access to customer models, datasets, training data, or inference traffic. It also noted that it does not store prompts or outputs by default [1].

Fireworks published an update on November 9 that widened the scope considerably. A third-party cloud provider hosting Fireworks R&D machines had suffered its own security incident on October 23. On October 24, the unauthorized party used that access to reach Fireworks development machines and run an unauthorized workload [2]. Fireworks reports that it detected this activity on the same day [2]. The resulting impact included resource metadata and billing information. For a limited set of accounts it also included secrets, datasets, and incomplete model weights [2]. Fireworks reported no impact to its Inference, Training SDK, or Nexus products, which it describes as ZDR by default. It also disabled self-serve dataset uploads and downloads for Batch Inference and dataset-based training while the investigation continued [2]. Fireworks' bulletins do not specify which intrusion reached the Notion dataset discussed below, but the October bulletin's statement that no datasets had been accessed points to the later intrusion.

Notion's account fills in the downstream effect. According to BleepingComputer's reporting, Fireworks alerted Notion on November 2 to an exposed dataset that initially appeared to contain only synthetic test data. By November 3, Notion had cut traffic to Fireworks, deleted its data from Fireworks systems, and halted training runs there. Cyber Security News reports that on November 6 Fireworks warned Notion that customer data might have been accessed [5]. On November 8, Fireworks confirmed that the dataset contained real Notion customer data [3]. SecurityWeek's reporting describes the November 2 notification as already concerning Notion AI data, so accounts of exactly what was known on that date differ slightly [4]. The exposed material consisted of Notion AI chat transcripts used for Insights, a feature that shows workspace administrators how their teams use AI. Notion stated that for most affected workspaces the transcripts were not attributable to a user or workspace, and that they did not include tool call results [3][4]. Notion disabled Insights, stopped serving open-weight models through

Fireworks, and advised affected customers, "We recommend treating this data as exposed." It also told them to rotate any passwords, API keys, or other credentials pasted into Notion AI during the relevant period [4][5].

Why this is a fourth-party problem

Most enterprise customers of a SaaS product evaluate that vendor as a third party. They review its security attestations, sign a data processing agreement, and accept its published subprocessor list. Notion's public AI security documentation gives a representative picture of what customers see at that level. It states that most models used for Enterprise workspaces offer ZDR. It states that a small number of models without ZDR are off by default and require owner opt-in. It states that for Free, Plus, and Business plans the AI provider retains data for 30 days or fewer. And it notes that some models are served through "inference providers" rather than by their original developers, without naming those providers on that page [6]. In CSA's assessment, that is a reasonably transparent posture by current market standards. Even so, customers would have had to follow the subprocessor list and then evaluate Fireworks' own controls to understand where their AI interactions were processed and in what form.

The term "fourth-party" describes this gap in visibility, not a legal category. Under the EU General Data Protection Regulation, a processor may not engage another processor without the controller's prior specific or general written authorization. It must also impose the same data protection obligations on that subprocessor by contract, and it remains fully liable to the controller for the subprocessor's performance [8]. The European Data Protection Board's Opinion 22/2024 went further. It indicated that controllers should be able to identify all processors in a processing chain at any given time, and that the controller's duty to verify sufficient guarantees extends to subprocessors further down the chain [9]. That expectation is stricter than much current market practice, in which customers typically see only the first tier of subprocessors. NIST SP 800-161 Rev. 1 likewise frames cybersecurity supply chain risk management as covering suppliers' own suppliers, not only direct vendors [10]. The Fireworks–Notion exposure tests these principles against a supply chain that has formed around AI features.

Security Analysis

Inference providers are a distinct subprocessor tier

The AI supply chain behind a typical SaaS feature has more layers than a conventional SaaS stack. A product team may call a frontier model provider directly. It may also route some workloads, especially open-weight models, through a specialized inference platform that hosts, serves, batches, and fine-tunes

models on rented GPU capacity. That platform may in turn run on one or more cloud or GPU hosting providers. Each layer is a plausible subprocessor, and each may receive prompts, outputs, embeddings, evaluation datasets, or fine-tuning corpora. The Fireworks incident shows that the risk at the inference layer goes beyond prompts in transit. Inference platforms also hold customer secrets for evaluation jobs, stored datasets, and partially trained weights [1][2], so they are concentrated stores of credentials and data from many downstream products.

The table below summarizes the parties involved in this incident as seen from a Notion enterprise customer. It is based on the public statements cited and does not assert facts beyond them.

Position relative to Notion customer	Party	Role in the incident	Customer's typical visibility
First party	Customer organization	Users entered prompts, sometimes including credentials, into Notion AI	Full
Third party	Notion	Routed open-weight model workloads and Insights processing to Fireworks [3] [4]	Direct contract, DPA, attestations
Fourth party	Fireworks AI	Held the Insights dataset that was accessed [2][3]	Subprocessor list entry, if reviewed
Fifth party	Unnamed cloud provider hosting Fireworks R&D machines	Its compromise on October 23 provided the route to Fireworks development machines [2]	Generally none

Zero data retention is a scoped claim, not a property of a vendor

A central detail in this incident is that ZDR commitments existed at two levels and data was still exposed. Notion has stated that Fireworks contractually committed to ZDR for Notion data [3]. Fireworks states that its Inference, Training SDK, and Nexus products are ZDR by default and showed no impact [2]. The public record does not fully explain how Notion's Insights transcripts came to be held in a Fireworks

dataset under those commitments. Notion's FAQ, as reported by BleepingComputer, states the ZDR commitment but does not explain how the Insights data came to be held in a Fireworks dataset [3]. One plausible reading of the available facts, which neither company has confirmed, is that the ZDR commitment was understood to cover real-time inference. Under that reading, the Insights classification and related training workflows used datasets stored at rest as part of a separate processing path. Notion's reported action of halting "training runs" at Fireworks [3] and Cyber Security News' description of "Insights classification jobs" [5] are consistent with this interpretation. They do not establish it.

Whatever the exact mechanism, the lesson for buyers holds. A ZDR claim attaches to a specific product, endpoint, or processing mode. It does not cover everything a vendor does with customer data. Analytics features, evaluation pipelines, batch jobs, quality review, and fine-tuning commonly need data to persist for some period. These are also workflows that are often added after an initial security review. When a vendor tells a customer that its AI provider uses ZDR, the customer should ask which workflows that statement covers. It should also ask whether any feature, including administrator-facing analytics, sends conversation data to the same provider through a path that retains it.

AI chat logs are credential stores

Notion's advice to rotate credentials pasted into Notion AI [4][5] reflects a practical reality: users debugging configuration files, troubleshooting API calls, or drafting scripts in an AI assistant can paste secrets into it. CSA has previously documented a related case in which AI model-gateway logs of coding-assistant sessions held GitHub credentials that were later used in a supply chain attack [17]. An AI chat transcript should therefore be classified as potentially credential-bearing data, whether it sits with the primary vendor or downstream. That classification affects how transcripts are retained, which subprocessors may receive them, and how incident response proceeds after an exposure. Notion's statement that most exposed transcripts were not attributable to a user or workspace [4] limits privacy harm in one dimension. It does not limit the value of a live API key appearing in the text. Secrets scanners and attackers typically find credentials by pattern, not by attribution.

Inference providers are plausible targets of the LLMjacking economy

Fireworks attributes the activity to a financially motivated cybercriminal group whose goal is to sell tokens obtained through unauthorized LLM access to illicit routing gateways [2]. That description fits the threat model GTIG and Mandiant published on November 2, 2026. Their report found that stolen cloud identity tokens were the dominant entry vector in 2026 LLMJacking incidents, with leaked LLM API keys also targeted. It also found that the stolen access was resold through "LLM mixers" that pool keys and proxy traffic, and that fraudulent charges at some victims exceeded US$100,000 in a single day and reached US$1 million cumulatively before access was revoked [7]. The GTIG report does not name Fireworks.

CSA's October 2026 analysis of Fireworks' first bulletin alongside the OpenCode and Railway disclosures treated these incidents as a campaign aimed at AI inference providers [16]. Together, the pattern suggests that inference platforms are attractive targets in two ways. Their own compute is directly monetizable, and they hold their customers' credentials for other services in evaluation secrets and environment variables [1].

For downstream enterprises, this changes the threat model for AI subprocessors. A compromised inference provider may expose data, and it may also expose credentials that let attackers spend the customer's money or reach the customer's other systems. Fireworks' first bulletin advised customers to rotate third-party credentials listed in their notices, noting that "removing a stored secret alone does not invalidate the underlying credential" [1]. That advice applies to any organization that has given secrets to an AI platform directly or through a vendor.

Disclosure lag compounds downstream

The timeline also shows how disclosure lag accumulates through a supply chain. Fireworks detected the October 24 intrusion on the day it occurred [2]. Notion received its first alert on November 2, nine days after detection, a period the public record does not explain. Fireworks warned Notion on November 6 that customer data might have been accessed [5] and confirmed real customer data on November 8, and Notion's customer communications and press coverage followed between November 10 and 12 [2][3][4][5]. Fireworks reported acting quickly to contain the intrusion, and Notion reported acting quickly once alerted. Even so, the end customer learned of the exposure more than two weeks after the intrusion, and only after the data's nature had been reassessed. Each additional layer adds investigation, attribution, and notification time. Customers' contracts and incident response plans should assume that fourth-party exposures will reach them late and incomplete.

Recommendations

Immediate Actions

Organizations that use Notion AI should check whether they or their workspace owners received an incident notice. If so, they should use Notion's security incident review to examine the affected transcripts and rotate any credentials that appear in them, following Notion's guidance [3][4]. Organizations that have stored secrets on Fireworks directly, or that received Fireworks notices, should rotate the identified credentials at the issuing provider rather than only deleting them from Fireworks, as Fireworks

recommends [1]. Beyond this incident, security teams should run a targeted secrets scan or a user advisory for AI assistants in general. Credentials pasted into any AI tool should be treated as candidates for rotation, and users should be pointed to approved secret-handling practices.

Third-party risk teams should also send a short, focused inquiry to their most data-intensive AI-enabled SaaS vendors. It should ask which inference providers process customer data for each AI feature, which features persist data at those providers for analytics, evaluation, batch, or training, and what retention applies to each. The answers should help show where the organization's exposure resembles the Notion case.

Short-Term Mitigations

Over the next one to two quarters, organizations should build AI subprocessor depth into vendor assessment. That means asking for the subprocessor chain behind each AI feature as well as the corporate subprocessor list, and recording it in the vendor inventory. This is consistent with the EDPB's view that controllers should be able to identify every processor in a chain [9]. Contracts and data processing agreements should define ZDR by processing path. They should require advance notice before any AI feature begins persisting customer content at a subprocessor, and they should require flow-down of breach notification timelines to the inference layer, consistent with GDPR Article 28's requirement that processors pass their data protection obligations down to subprocessors by contract [8].

Internally, organizations should classify AI conversation logs as sensitive data that may contain credentials, and apply retention limits and access controls to match. Where vendors offer administrative controls, organizations should review whether AI analytics and usage-insight features are enabled and whether their value justifies the additional data flow. Notion's documentation notes that models without ZDR are off by default and require owner opt-in [6]. Controls like these let customers reduce fourth-party exposure without waiting for contract renegotiation.

Strategic Considerations

Over the longer term, the inference layer of the AI supply chain needs assurance in its own right. Specialized inference and fine-tuning platforms can hold customer data, credentials, and model artifacts from many downstream products, which can make them shared points of failure. Buyers and vendors should expect, and increasingly request, AI-specific attestations from these platforms. Assessors who evaluate SaaS vendors should treat the vendor's oversight of its inference providers as a control to test, not a question to note.

Organizations building AI features on open-weight models should also consider the trade-off between outsourced inference and self-hosting. Self-hosting removes a subprocessor but takes on operational and security burden, and a specialized provider may run tighter controls than an in-house deployment would, so it is not automatically safer. Either way, the decision should account for the data each workflow sends out, not only cost and latency. Finally, incident response playbooks should include a fourth-party AI exposure scenario. The scenario should assume late, evolving disclosures, uncertain data attribution, and a need for credential rotation across systems the AI assistant never directly touched.

CSA Resource Alignment

CSA's AI Provider Concentration Risk: Enterprise Resilience [11] argues that enterprises should manage AI providers as third-party risk relationships rather than procurement decisions. It recommends AI dependency inventories, contractual protections, and governance visibility over the AI stack. The Fireworks–Notion exposure adds a confidentiality dimension to that resilience argument. The dependency inventory the paper recommends should extend past the AI features a vendor ships to the inference providers that serve them, because an exposure at that layer surfaces as the vendor's incident.

CSA's research note The Agents You Didn't Choose: Embedded AI and Liability [12] describes how organizations inherit AI capabilities through SaaS renewals and platform updates, often without visibility into what data those capabilities touch. Notion's Insights feature is an example of that pattern at the data-flow level. It was an administrator-facing analytics capability whose processing path ran through a provider the customer did not choose. That note's recommendation that inventories include AI embedded in vendor products applies directly to inference subprocessors. CSA's research note Inference Providers as the Prize: Platform-Chain Campaigns Against AI Serving [16] examined the earlier phase of the Fireworks incident as part of a broader campaign against AI serving platforms, and its analysis provides the threat context for the subprocessor exposure described here.

On assurance, CSA's STAR for AI program [14] gives buyers a structured way to request AI-specific security assurance from vendors. CSA's research note Who Assures the Assurers? Third-Party AI Assessment Standards [13] cautions that a certificate is only as reliable as the body that issued it and the scope it covers. Read together, they suggest that buyers should check whether a vendor's AI assurance scope covers its inference subprocessors and data-persisting AI workflows. A ZDR statement, like a certificate, should be read for what it covers.

Finally, the AI Controls Matrix (AICM) v1.1 [15] provides control language for the recommendations above. Its Supply Chain Management, Transparency and Accountability domain covers subprocessor inventory and flow-down. Its Data Security and Privacy Lifecycle Management domain covers retention by

processing path and classification of AI conversation data. Its Security Incident Management, E-Discovery and Cloud Forensics domain covers notification obligations across the supply chain, and its Identity and Access Management domain covers the handling of secrets entrusted to AI platforms.

References

[1] Fireworks AI. "October 2026 Security Incident." Fireworks AI Security Bulletin, October 9, 2026 (updated October 10, 2026).

[2] Fireworks AI. "October 2026 Security Incident – Update." Fireworks AI Security Bulletin, November 9, 2026.

[3] Bill Toulas. "Notion says Fireworks breach exposed customer AI chat transcripts." BleepingComputer, November 11, 2026.

[4] Ionut Arghire. "Notion Customer Data Compromised in Fireworks AI Hack." SecurityWeek, November 12, 2026.

[5] Guru Baran. "Notion Confirms AI Chat Transcripts Exposed in Fireworks AI Breach, Affecting Customer Workspaces." Cyber Security News, November 10, 2026.

[6] Notion. "Notion AI security & privacy practices." Notion Help Center, accessed November 12, 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] European Union. "Art. 28 GDPR – Processor." General Data Protection Regulation (EU) 2016/679, 2016.

[9] European Data Protection Board. "Opinion 22/2024 on certain obligations following from the reliance on processor(s) and sub-processor(s)." EDPB, October 2024.

[10] NIST. "SP 800-161 Rev. 1, Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations." National Institute of Standards and Technology, May 2022 (updated November 2024).

[11] Cloud Security Alliance. "AI Provider Concentration Risk: Enterprise Resilience." CSA AI Safety Initiative, June 2026.

[12] Cloud Security Alliance. "The Agents You Didn't Choose: Embedded AI and Liability." CSA AI Safety Initiative, October 2026.

[13] Cloud Security Alliance. "Who Assures the Assurers? Third-Party AI Assessment Standards." CSA AI Safety Initiative, September 2026.

[14] Cloud Security Alliance. "STAR for AI: AI Security Assurance Program." Cloud Security Alliance, accessed November 12, 2026.

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

[16] Cloud Security Alliance. "Inference Providers as the Prize: Platform-Chain Campaigns Against AI Serving." CSA AI Safety Initiative, October 2026.

[17] Cloud Security Alliance. "LiteLLM 1.107.0: When Model-Gateway Logs Become the Credential Store." CSA AI Safety Initiative, October 2026.