STRATEGIC INTELLIGENCE DIVISION

The breach arrives as an invoice, and the platform cannot pay it

Twelve AI developer platforms disclosed credential thefts between 4 October and 1 November 2026. The key was the means, not the end: two enterprise keys ran about 800,000 and 1.7 million Bedrock calls, and the charges landed on the victims' own cloud accounts. Zeabur, which stored its customers' keys, says it cannot pay.

Annika Sorensen, Strategic Intelligence Division · 4 November 2026 · 11 minute read

In brief

- Two enterprise keys ran about 800,000 Bedrock calls in 30 hours and about 1.7 million in 13 hours, and at list price those two cases alone likely ran up over $100,000 in usage.
- Only about 6% of the first case's calls were throttled, because the key's policy allowed cross-region inference profiles and the load spread across regions.
- Mandiant found LLM services switched off at the time of compromise: the attacker enabled the API and generated the bill in minutes or hours, so the invoice arrives for a product the victim never turned on.
- A single-token request validates a stolen key without the volume that endpoint detection and SIEM anomaly rules watch for, so any alert arrives after the spend.
- Zeabur says it cannot pay the third-party charges its exposure created and ceases operations on 31 December 2026, which is 57 days from 4 November 2026.

> *“Under our current financial circumstances, we are unable to pay for third-party service charges arising from the October 2026 security incident.”*

Zeabur, follow-up on the October 2026 security incident, 28 October 2026

Zeabur held its customers' API keys and other credentials on the platform, exposed them when its systems were breached on 14 October 2026, and has now told those customers it cannot pay the third-party charges the exposure created. What it offered instead was an attachment: its incident investigation report, to be sent with a fee-waiver request to each provider. Mandiant's 2026 cases include fraudulent charges exceeding $100,000 in a single day and cumulative charges reaching $1 million at one organization before access was revoked, and the meter that records all of it is the victim's own cloud account. The decision this forces is not a security programme. It is key custody and a spend ceiling, signed by the CISO with the CFO, and the date that makes it urgent is 31 December 2026, when Zeabur's clusters stop.

The loss arrives as consumption, so nobody at the victim is looking for it

Wiz traced two enterprise keys from a vendor breach into Bedrock usage that, at list price, likely ran up over $100,000. Case A made approximately 800,000 calls in 30 hours and about 90% of them succeeded. Case B made approximately 1.7 million calls in 13 hours from a single Level 3 address, with over a million successful. Wiz describes both token profiles as resembling resold coding-agent traffic: a handful of uncached input tokens in, a few hundred tokens out, and a long tail of calls carrying tens of thousands of uncached input tokens.

The attribution problem is structural rather than a logging gap. In the GCP intrusions Wiz CIRT investigated, the attacker linked external, attacker-controlled projects to the victim's billing account, so “prompt contents and model outputs remained completely hidden from the victim organization” while the victim carried financial responsibility for the consumption. The attacker also demoted the incumbent

billing administrators from their administrative roles down to read-only billing viewers, which removed their ability to detach the rogue projects from the console. The event then sits in two registers at once, a cost loss and a data event that cannot be reconstructed, and neither owner can close it on their own.

The evidence for a dispute sits with a company switching itself off

Zeabur disables dedicated server renewals on 1 December 2026 and shuts all dedicated servers and shared clusters by 31 December 2026. A customer who wants a provider to waive the charges needs one thing from Zeabur: the incident investigation report it published for exactly that purpose. A customer who wants its data back needs three more, the data deletion date, the final data export deadline and the date the support desk closes, and those three are, in Zeabur's own announcement, “To be announced”. A dispute file assembled after the support desk closes has no author. Source: Zeabur, “Zeabur is shutting down: what happens next”, 17 October 2026, and “Zeabur follow-up on the October 2026 security incident”, 28 October 2026.

The attacker chose the routing that regional quotas cannot bound

Case A's key permitted global.* cross-region inference profiles, and the actor routed every call through them. Because global routing spread the load across regions, only about 6% of calls were throttled, which is why a rate limit and a volume alert are the wrong instruments against this. Where the SIEM does watch volume, Mandiant describes the counter: automated credential validation with minimal single-token requests (max_tokens=1), which confirmed the viability of stolen keys while remaining undetected by the victim's endpoint detection and response and its SIEM volume anomaly alerts. Case B shows the control holding briefly and the attacker working around it: that key allowed only us.* profiles, thousands of calls failed over four hours, and once a checker found the working model identifiers, 1.7 million calls went through one address in 13 hours. GTIG records the step above the key as well: a follow-on cluster took pre-validated Virtual Keys from the actor that established access and exploited a user ID rebind vulnerability in the /key/regenerate endpoint of the victim's LiteLLM gateway (CVE-2026-40217), raising standard Virtual Keys into full PROXY_ADMIN master administrative credentials.

Twelve platforms in 28 days is a category failure, not twelve vendor failures

Twelve platforms, 4 October to 1 November 2026, and what each has said

Platform	Disclosure	What the vendor has said
Fireworks AI	6 October 2026	Attributes the incident to credential access
Andromeda	9 October 2026	Calls it part of an “industry-wide attack”
Zeabur	14 October 2026	Credentials stored on the platform were exposed; “we are unable to pay for third-party service charges”; shutting down
Lightning AI	15 October 2026	Attributes the incident to credential access
OpenCode	16 to 17 October 2026	A “legacy system”; attributed to an attempt at distillation
Railway	20 October 2026	Initial access via “privilege escalation in our build infrastructure”
Warp	23 October 2026	A third-party vendor dependency
CommandCode	28 October 2026	Attributed to an attempt at distillation
Chainstack	29 to 31 October 2026	No TTPs or IOCs shared
Northflank	30 October 2026	“Please rotate all user defined credentials”; the same page states no breach since 1 April 2019
Trigger.dev	1 November 2026	The attacker gained access to environment variables for a subset of organizations
Brev (NVIDIA)	4 to 20 October 2026, reported later	No TTPs or IOCs shared

Source: Wiz, “Black October for AI Developer Platforms”, 30 October 2026, updated 2 and 3 November 2026, for the disclosure dates and the vendor statements it records, including its characterisation of the statements made by Fireworks AI, Andromeda, Lightning AI, OpenCode, Railway, Warp and CommandCode, which are Wiz’s account of those statements rather than those vendors’ own published text; Zeabur blog, 17 and 28 October 2026; Northflank security page, read 4 November 2026; Trigger.dev incident report, 1 November 2026. Mosaic Singularity analysis, November 2026.

Wiz states plainly that there is “no single vulnerability or dependency linking these attacks”. That sentence is what makes the category reading correct and the single-vendor reading wrong: twelve different root causes, one objective, and a market that pays for the result. A buyer who asks each supplier whether it has been breached gets twelve true answers and no exposure picture, because in the reports that describe the mechanism, the same thing failed: a platform stores its customers' provider credentials in its own environment.

The resale layer holds the only margin, and revocation is what ends it

Mandiant puts a price on the demand side. Jet Night, a token routing service powered by stolen third-party keys, executed more than 500,000 LLM API calls in less than eight hours and passed nearly one million dollars in token costs to the victim organization. It sells an unlimited plan at approximately $60 per month with alleged access to frontier models, a service level GTIG says would typically run between $100 and $250 per month. GTIG believes the service relies on stolen keys to operate profitably, and it notes the multiple outages the service has advertised on its public Telegram channel, which it says is likely due to stolen keys being revoked as victims discover their compromises. That is the whole lifespan of the advantage: it survives until the victim reads the bill.

The supplier's own security page documents the long-lived secret

Northflank's security page carries a banner at the top: “Please rotate all user defined credentials.” Further down the same page it describes the integration model that creates the exposure. On authorisation by API

key, “The user provides API keys to their cloud account which are encrypted and stored in our database. The secrets are long-lived.” The alternative, in which the customer grants a Northflank-controlled IAM principal access to its cloud account, is published there as “the recommended authorization strategy for cloud provider integrations”, and “With this approach, Northflank does not need to store long-lived tokens at all.” The mitigation for secrets held inside the platform, external secret managers, is listed as “currently in beta”.

Trigger.dev's disclosure is one sentence about the same asset: “The attacker gained unauthorized access to environment variables for these organizations.” What the two pages have in common is that the question, will you store our key and for how long, has a written answer from the vendor after the incident. It is the question worth asking while the answer is still a term you can negotiate.

What each seat owes, and the date it falls due

- CFO: re-cut the AI opex line by key and by project, and require that consumption which cannot be attributed to a person, a project or a product be reported as an incident cost rather than as business demand. The evidence for the requirement is two keys and about 2.5 million calls no internal owner ordered. By 31 December 2026.

- CISO: retire long-lived provider keys from application and vendor environments in favour of short-lived brokered credentials, and set a spend ceiling per key and per project. Mandiant's guidance is to use spending caps and granular API quotas, because the volume the anomaly rules watch is the volume the attacker avoids: single-token validation first, then a burst routed across regions. Extend patch management to the model gateway itself, because GTIG's guidance names gateway patch management and cites CVE-2026-40217, used to raise standard Virtual Keys to PROXY_ADMIN. By 31 December 2026.

- General Counsel: read the two clauses that decide who pays, meaning whether the supplier bears third-party charges arising from its own exposure, and what it must hand you to support a dispute. Zeabur's answer to the first is that it cannot pay and that customers should seek fee waivers from their providers, so only the second clause has value after the fact. Obtain Zeabur's incident investigation report and a copy of the charge evidence by 31 December 2026, when its dedicated servers and shared clusters stop, and do not wait for its support desk, which closes on a date it has not announced.

- CIO: produce the list, naming which of the twelve platforms sit in the estate, which of them hold a long-lived provider key, and on what rotation and notification terms. Northflank publishes the alternative configuration in writing, so the question has a yes or no answer per supplier rather than a maturity score. By 31 December 2026.

The board's decision is a key-custody register and a spend ceiling, due 31 December 2026

The decision that reaches the board is not whether to fund a security programme. It is whether to require the CFO and the CISO to bring two artefacts to the board by 31 December 2026: a register of every supplier holding a long-lived provider key of ours, and a spend ceiling per key and per project with the name of who can raise it. The trigger for a report back is the first month in which AI consumption on any cloud account cannot be tied to a person, a project or a product, or any single key crosses its ceiling, and either event brings the register to the board with the invoice beside it. The questions below are how the board tests whether those two artefacts exist.

- Which services appear on the AI bill that we never enabled, and who would catch the next one being switched on?
- What share of the October 2026 inference spend can be tied to a person, a project or a customer, and who signed for the rest?
- Which supplier holds a long-lived key of ours, for how long, and what does the contract say about the charges it causes?
- What is the spend ceiling on each key, who can raise it, and does it alert before the ceiling is reached or after?
- If a supplier holding our keys stopped trading without notice, what would we have to do ourselves, and how many days would it take?

What to watch, and when it resolves

Two dates are fixed. On 1 December 2026 Zeabur disables dedicated server renewals and pauses services on shared clusters. On 31 December 2026 its dedicated servers and shared clusters stop, and any fee-waiver file that depends on its incident report has to be assembled before a support desk closes on a date still unannounced. The variable to watch is whether any provider publishes a standing reversal of these charges, because that single document would move the loss back up the chain.

The reading here is that the charge stays with the buyer, and that no upstream party has committed to absorbing it. It fails on either of two events: a cloud or model provider publishing a standing reversal of LLMjacking charges, or one of the twelve stating that it has funded its customers' provider charges. Expect the amount rather than the incident to be disclosed first, because billing is the only part of this that leaves a number: watch the Q1 filings before 31 March 2027, where a charge of this size, if it is material, would have to appear in a contingency note.

Seven documents carry this, and four of them are the platforms' own words

- Google Threat Intelligence Group (Mandiant), The Rise of LLMJacking: How Threat Actors Monetize Compromised Cloud and Model Resources, 2 November 2026.
- Wiz, Black October for AI Developer Platforms, 30 October 2026, updated 2 and 3 November 2026.
- Wiz CIRT, From Phishing to LLMjacking, 20 October 2026.
- Zeabur, Zeabur follow-up on the October 2026 security incident, 28 October 2026.
- Zeabur, Zeabur is shutting down: what happens next, 17 October 2026.
- Northflank, security page, undated, as quoted on 4 November 2026.
- Trigger.dev, Security Disclosure: Sunday November 1st, 1 November 2026.

Sources

- Zeabur, follow-up on the October 2026 security incident, 28 October 2026
- Wiz, “Black October for AI Developer Platforms”, 30 October 2026, updated 2 and 3 November 2026, for the disclosure dates and the vendor statements it records, including its characterisation of the statements made by Fireworks AI, Andromeda, Lightning AI, OpenCode, Railway, Warp and CommandCode, which are Wiz’s account of those statements rather than those vendors’ own published text; Zeabur blog, 17 and 28 October 2026; Northflank security page, read 4 November 2026; Trigger.dev incident report, 1 November 2026. Mosaic Singularity analysis, November 2026.