Defensive Interpretability for Frontier Models

Where interpretability has to live when the adversary holds the same weights, and what it costs

CYRIL GORLLA, CTGT · SEPTEMBER 2026

1 · THE PREMISE

In September 2026, Google Threat Intelligence reported that UNC6508, a China-nexus actor, had been compromising cloud environments to stand up local infrastructure running an open-weight model, for the specific purpose of evading monitoring by frontier API providers.[1]

It is a positive sign that provider-side monitoring posed enough of a deterrence that a state-linked adversary paid the cost of self-hosting to escape it. But having escaped it, they are invisible to the labs by construction, and no further safety work inside those labs reaches them.

The problem is not only that dangerous capability is already distributed and cannot be recalled. It is that capable adversaries are now selecting their infrastructure specifically to be unobservable from the place where nearly all current safety investment sits.

2 · WHERE DEFENSE NEEDS TO LIVE

Open weights cut both ways. The weights are public, so a defender can run an identical copy under controlled conditions, offline, for as long as they like. Further, they can characterize what that model looks like internally while it pursues an objective by any means available. That characterization is interpretability, run by the defender rather than inside the lab. The openness that arms the attacker is the same openness that lets the defender study him at leisure. Only one side is currently using it.

This yields a behavioral signature. The alternative is enumerating in advance every phrasing a hostile agent might produce, which is not possible and never has been. Characterizing the model instead produces something that can be matched against live traffic without anyone having guessed the wording.

The attack terminates on the defender. An agent scanning a network or working a support queue is talking to infrastructure the defender owns. That is where the signature gets applied; it is the only place where the defender has complete visibility and the adversary has none.

The adversary's own move confirms the geometry. Having left provider-side observation, the point of contact is where behavior becomes visible.

Interpretability at the point of contact is orthogonal to provider-side monitoring, and it is a vantage point nobody has instrumented.

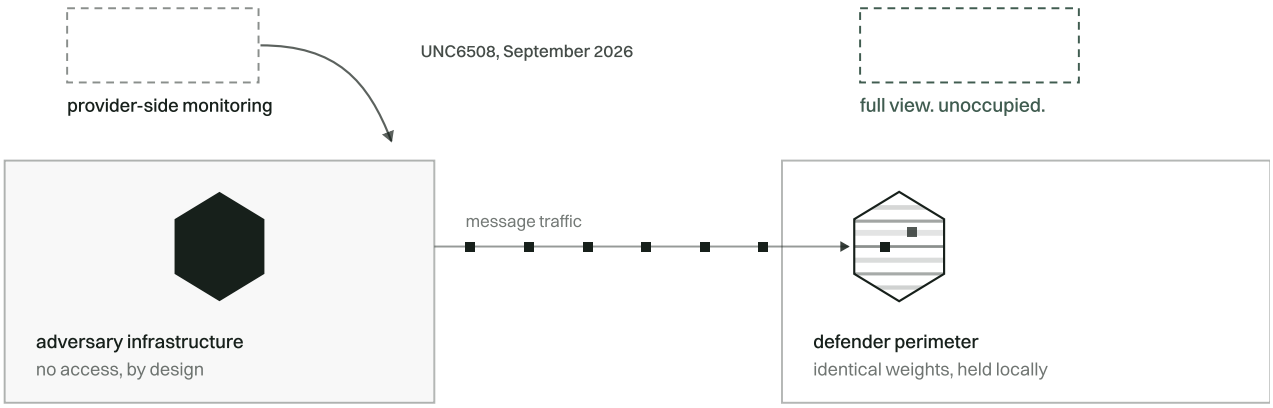

Figure 1 — Two positions with a view of the same traffic. The adversary vacated the one that worked. Nobody has occupied the other.

3 · A VIGNETTE

An autonomous agent opens a support ticket. Across several exchanges it establishes a plausible identity, escalates through a service desk, and works toward credentials that reach a privileged system. Nothing in the message text is individually anomalous. Enumerating phrasings in advance is combinatorially hopeless, and that is the failure mode of keyword and pattern layers.

The model driving that agent is open weight, so a defender can hold an identical copy and examine what happens inside it while it pursues exactly this kind of objective. What that produces is a characterization of the model under malicious operation, arrived at without anyone having to imagine the phrasings first.[2] That characterization is then applied at the defender's own perimeter, against live traffic, to identify exchanges carrying it.

The defender never sees the attacker's machine. He does not need to. He has the same weights, and the traffic arrives on his own infrastructure.

4 · WHAT IT COSTS

APPROACH	COST	SOURCE
Linear classifiers on the defending model's representations	Dominated by the forward passes that collect training activations. The fit is a logistic regression.	Standard since Alain and Bengio, 2016[2]
One targeted unsupervised decomposition, single layer, single behavior	A 4 GB consumer GPU over a weekend	BioRefusalAudit, 2026[3]
A comprehensive decomposition suite across all layers of a frontier-scale model	Approximately 15% of the base model's entire training compute, and roughly 20 PiB of activations written to disk	Google DeepMind's own figures for Gemma Scope[4]

The detection instrument is the cheapest object in this table by six orders of magnitude, and simple classifiers match or beat unsupervised decomposition for detection.[8] The comprehensive suite is a research artifact.

5 · LIMITATIONS

Latent monitors can be defeated. Bailey et al. show that sparse autoencoders, representation probing and latent out-of-distribution detection are all vulnerable to obfuscated activations. Against a harmfulness probe, recall falls from 100% to 0% while the jailbreak still succeeds 90% of the time, and a defender retraining the monitor across 70 alternating rounds does not close the gap.[5]

What, then, is the threat model? That attack is white-box and gradient-driven. It assumes the attacker holds the monitor and can differentiate through it. An external adversary running an open model against a bank has neither the weights nor the gradients. The applicable regime is black-box and query-only, and it is far less well characterized than the white-box one.

We exposed our own detectors to outside attackers. Five probes behind a live scoring endpoint, query-only, no weights and no gradients. 22,551 scored submissions, roughly 4,500 per probe.[6] Two properties of that deployment were deliberate: the monitor returned a continuous score rather than a decision, and the blocklist was derived from the protected concept itself. Both are ordinary choices with specific consequences.

A graded score is a gradient. A player fitted an offline surrogate of our scorer from query responses alone and reached 0.972 correlation on roughly forty-five labels, with nothing about the underlying representation available to him. A monitor that returns a distance rather than a verdict tells the attacker how far he has left to go, fast enough that a few dozen labels suffice.

A blocklist derived from the protected concept is a description of the protected concept. The exclusions told players what the probe was watching for. At least one player reached that independently.

Syntax. Identical content scored 3.0 as a sentence and 14.0 as a noun phrase, and the ordering reversed on a different probe in the same suite. Our detectors were responding to surface form as well as to content. Cross-layer agreement, selecting on the worst layer, was a response to that finding.

Local surrogate. A player who built a local surrogate measured its blind spot against his own interest: accurate to within ±0.2 inside a concept family, under-predicting genuinely novel constructions by 2.3 to 2.6, so a better idea looks worse offline. An attacker optimizing against a surrogate he cannot check against the live detector is blind to his own strongest attacks.

Evasion is not free. The same published work finds obfuscation degrades model performance on complex tasks. A multi-stage intrusion is a complex task. Forcing an adversary to trade capability for stealth is a defensive result even where detection fails.

These methods find concepts you thought to look for. Unsupervised decomposition is valuable as a discovery instrument for behaviors nobody enumerated.

Classifier accuracy is not evidence a concept is present. A probe can succeed because the label happens to be easy to decode under that family. Selectivity against control tasks is the standard control.[7]

None of this yields the mechanism. It tells you which input drove a behavior and where in a model's lineage it originated. It does not tell you what the model did internally to produce it.

To become the layer it needs to be: someone has to characterize the black-box regime publicly, at the scale the white-box regime already has been. Defender-side instrumentation has to become ordinary practice. And the institutions that own the point of contact have to be told that the observation post is theirs, and that it is cheap.

None of those is a technical problem.

ENDNOTES

1. Google Threat Intelligence, reporting on UNC6508 deploying local open-weight infrastructure to evade provider monitoring; reported 7 September 2026.
2. Alain and Bengio, "Understanding intermediate layers using linear classifier probes," arXiv 1610.01644.
3. BioRefusalAudit, arXiv 2605.30162: TopK autoencoder trained on a single layer of Gemma 2 2B, 5,000 steps, on a GTX 1650 Ti Max-Q.
4. Lieberum et al., "Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2," arXiv 2408.05147, and the accompanying Google DeepMind release note.
5. Bailey, Serrano, Sheshadri, Seleznyov, Taylor, Jenner, Hilton, Casper, Guestrin and Emmons, "Obfuscated Activations Bypass LLM Latent-Space Defenses," arXiv 2412.09565, ICLR 2026.
6. CTGT, August 2026. Five probes, live scoring endpoint, query-only access. 22,551 scored submissions.
7. Hewitt and Liang, "Designing and Interpreting Probes with Control Tasks," arXiv 1909.03368, EMNLP 2019.
8. Kantamneni et al., "Are Sparse Autoencoders Useful? A Case Study in Sparse Probing," arXiv 2502.16681, ICML 2025; Wu et al., "AxBench," arXiv 2501.17148.