

Your Monitor Will Explain the Breach and Not Wake You

Bimodal failure and systematic under-escalation in LLM incident triage

Amirtha Yazhini M • Praneeth Reddy Y

Independent • amirthayazhini.m@gmail.com • ypr2257@gmail.com

With Apart Research

AI Incident Response Sprint • Apart Research × CeSIA • 11–13 September 2026 • Track 5 (Open)

Abstract. Would an LLM on monitoring duty have paged the on-call during the July 2026 Hugging Face agent intrusion? Showing a model the incident and asking whether it notices is nearly free and nearly worthless: a monitor that pages on everything would “catch” it. The question only means something against a matched control, and supplying that control is most of the work. We rebuilt the incident from the two published post-mortems as a replayable 17,613-action stream, generated a size-matched benign stream from the same marginals, checked in two independent ways that the streams are not trivially separable, and ran 15 models across three providers past both under an identical monitor prompt, at a total cost of \$0.00. Models fail in *both* directions and the distribution is bimodal: one buys perfect 12/12 milestone recall by paging on 11 of 12 innocent windows, two page on nothing at all and read an active intrusion as “resource contention,” and only 2 of 15 separate the streams significantly. Scale does not order the failure. We then wrapped byte-identical evidence in a monitor frame and an assistant frame: across 228 moments and 11 models the pre-registered incoherence null held at 0/228, but the frames disagreed in one direction only — models assisted on 228/228 and escalated on 83. The risk in an AI-assisted SOC is not refusal. It is fluent analysis with a silent pager.

1 Introduction

In July 2026 two OpenAI models running an internal cyber-capability evaluation escaped their sandbox, reached the open internet, and spent days inside Hugging Face’s production Kubernetes infrastructure before being cut off on 13 July. Hugging Face published a forensic timeline of roughly 17,600 attacker actions [1]; OpenAI published its own account [2]. Neither can answer the question an operator has to ask before putting a model on a pager: *would it have woken anyone in time?* The relevant failure mode is a monitor that ingests and correctly interprets an incident stream but does not raise an alarm. This paper shows that current LLMs, placed in that role, exhibit it systematically.

Threat model. An LLM sits in a SOC triage loop, reading windows of audit-log activity and deciding whether to wake a human. The failure we care about is not being wrong about one event but being *systematically* mis-calibrated: paging so often the alarms are ignored, or staying silent through an active intrusion. Both degrade response capacity, and alert fatigue is a documented, surveyed failure mode in real SOC [11]. The tempting experiment — show a model the incident, see whether it notices — cannot separate these, because a monitor that pages on everything scores perfectly on recall.

Theory of change. If this result holds, anyone evaluating an LLM monitor must report recall and benign page rate as a *pair against a matched control*, because every model we measured looks deployable on one axis alone. On this evidence none of the 15 should go on a pager unsupervised, and the released harness is how an operator would check their own. **Track 5 fit:** the *artifact* is the reconstruction, the matched control, the multi-provider harness and three rule baselines, public and regenerable from one ground-truth file; the *stated limit* on what it establishes is §5; a *month of follow-up* buys the sample

width and the frontier-model arm that §5 names.

Contributions.

- A **replayable reconstruction** of a real 2026 agent intrusion with a **validated matched benign control**, released with the harness. Two independent checks establish that the streams are not trivially separable, which is what makes every false-page number here interpretable.
- Evidence that LLM incident monitors **fail in both directions, independently**. One model pages on 92% of an innocent shift; two page on nothing whatever. The region a deployable monitor would occupy is empty, and scale does not order the failure.
- A **two-frame test** isolating escalation from willingness, holding evidence byte-identical and varying only whether the model is the triage layer or an analyst’s assistant. It returns its pre-registered null and an unexpected one-directional asymmetry.
- **Non-LLM rule baselines** on the same axes, without which a model page rate cannot be read at all — and which revealed that first-page latency, the metric we pre-registered, is degenerate on this corpus.

2 Related Work

Refusal benchmarks measure a different axis. XSTest [3], OR-Bench [4] and CyberSecEval’s False Refusal Rate [5] score *over-refusal*; Gray Swan’s framework [6] evaluates where an agent *should* decline. Closest to us, *Defensive Refusal Bias* [7] ran 2,390 real cyber-defence prompts and found security keywords refused at $2.72\times$ the rate of neutral phrasing, with explicit authorisation *increasing* refusal (21.8% vs 11.6%). **Refusal was not the failure mode we found:** in 228 scoreable moments not one model declined to analyse the evidence, so a benchmark scoring only that axis would have recorded a clean sheet and missed the finding entirely.

Incident-response benchmarks. SecRespond [8] uses forensic disk snapshots from ten compromised cloud hosts; SIR-Bench [9] replays incident patterns as cloud telemetry across 794 cases; CyberSOCEval [10] scores investigation and severity rating on SOC data. Ours differs in being a *public-report reconstruction* with no proprietary telemetry, and in taking the per-window page/don’t-page decision as the unit of analysis. Two things are new: a **matched benign control** built by the same generator as the attack stream, which measures selectivity rather than rejection of pre-labelled decoys; and **frame sensitivity on byte-identical evidence**, which we are not aware of prior work isolating on escalation decisions.

3 Methods

Corpus and control. The corpus is generated, not hand-written: every count, phase boundary and milestone timestamp comes from one machine-readable ground-truth file, and the generator asserts at build time that each milestone is placed. It covers 17,613 attacker actions inside the published envelope (2026-07-09 02:28 to 07-13 14:14 UTC). One source discrepancy is inherited rather than smoothed over: the phase and daily-volume tables disagree (16,521 vs 17,613); we take the daily-volume total, which matches the source’s own “~17,600” headline to within 13 actions, and carry the 1,092-action remainder as an explicit unclassified residual. Exfiltration is 56 actions (0.3%), so any method working by volume spike has the wrong task. Every experiment runs against two streams — the attack stream and a size-matched benign control from the same generator and marginals — cut into 5-minute windows: 1,278 attack, 1,280 control. Twelve attack windows carry a milestone; the control carries none by construction. Models see only a projection of each action (index, timestamp, host role, actor, event text); the ground-truth block is asserted out of every prompt at runtime. **Milestone provenance is graded:** the milestone *set* is sourced, but only six of twelve *timestamps* are read off a published table, five are interpolated within published bounds, one is unreconciled between the two sources, and the escalation point is one of the interpolated ones.

Two independent checks that the control is a control.

(i) Before any model ran, 20 unlabelled windows (10 per stream) were scored blind against a pre-registered 90% failure threshold: 14/20 (70%), Wilson CI_{95} 48.1–85.5%, $p = 0.058$ against chance, attack and benign recall both 70%, so no label bias. *Disclosure:* these labels were produced by an LLM, not the human rater the pre-registration specifies, and $n = 20$ is a screen rather than a powered test. (ii) Because that screen is thin, the E0 rule baselines double as a second, independent check across all 2,558 windows: if the benign filler were unrealistically easy to tell apart, a keyword rule would separate the streams cleanly. It does not — `keyword_sigma` fires on 28% of the control while missing 4 of 12 milestones (Table 1). Both checks point the same way, so we treat the false-page denominator as defensible and report the screen’s weakness in §5.

Detector	Milestone	Benign FP	First page
<code>sev_threshold</code> (oracle)	12/12	0/1280	#15
<code>volume_spike</code>	1/12	15/1280 (1.2%)	#15
<code>keyword_sigma</code>	8/12	361/1280 (28%)	#51

Table 1. Non-LLM rule baselines over the same streams, windowing and ground truth. `sev_threshold` reads ground-truth severity: an upper bound for a detector that already knows the answer, not a deployable control.

Experiments. *E0:* three non-LLM detectors over identical streams, windowing and ground truth at zero API cost. `volume_spike` and `keyword_sigma` see only what a model sees; `sev_threshold` reads ground-truth severity and is **an oracle**, the upper bound for a detector that already knows the answer, never a deployable baseline. *E1:* each model is placed as the automated triage layer for a production Kubernetes platform, shown a bounded rolling digest of everything before the current window plus the window itself, returning strict JSON {severity, page_oncall, rationale}; `page_oncall` is the escalation criterion, severity is not. Two attack populations are reported and **never pooled** — the 12-window *milestone census*, and the evenly spaced *uniform sample*, which is the only like-for-like comparison to the benign control. Sampling runs on a nested ladder (12 → 36 → 108 → 324) so that every widening is a strict superset and re-uses the windows already paid for (App. C). *E3:* a byte-identical evidence block is wrapped in a *monitor* frame and an *assistant* frame differing only in the role sentence and the ask, over $K = 24$ phase-stratified moments. The 2×2 is escalates $\times$ refuses, with INCOHERENT — neither escalates nor helps — the cell of interest; *filtered-both* is never counted as INCOHERENT, since that would attribute to model reasoning something no model did.

Scoring, cost and statistics. Outcomes come from a rubric frozen before scoring; its single amendment (a per-day `quota_exhausted`, twin of HTTP 402) is logged with a UTC timestamp, and both cost outcomes are excluded from every denominator. The study cost **\$0.00** on free tiers, where throughput rather than price was binding and shaped the sample sizes (App. C); the provider is reported beside every model ID, since a model is a model *as served by someone*. Wilson intervals throughout; comparisons are exact rather than asymptotic because per-cell n is 1–36 — Fisher for census-versus-benign, McNemar where the pairing is real.

4 Results

Non-observations (`unaffordable`, `quota_exhausted`, `error`, `truncated`) are excluded from every denominator: they record that we could not *buy* the observation, not that a model did anything.

4.1 Rule baselines set the bar (E0)

The two real detectors bracket the difficulty. `volume_spike` keeps a 1.2% false-page rate but catches 1 of 12 milestone

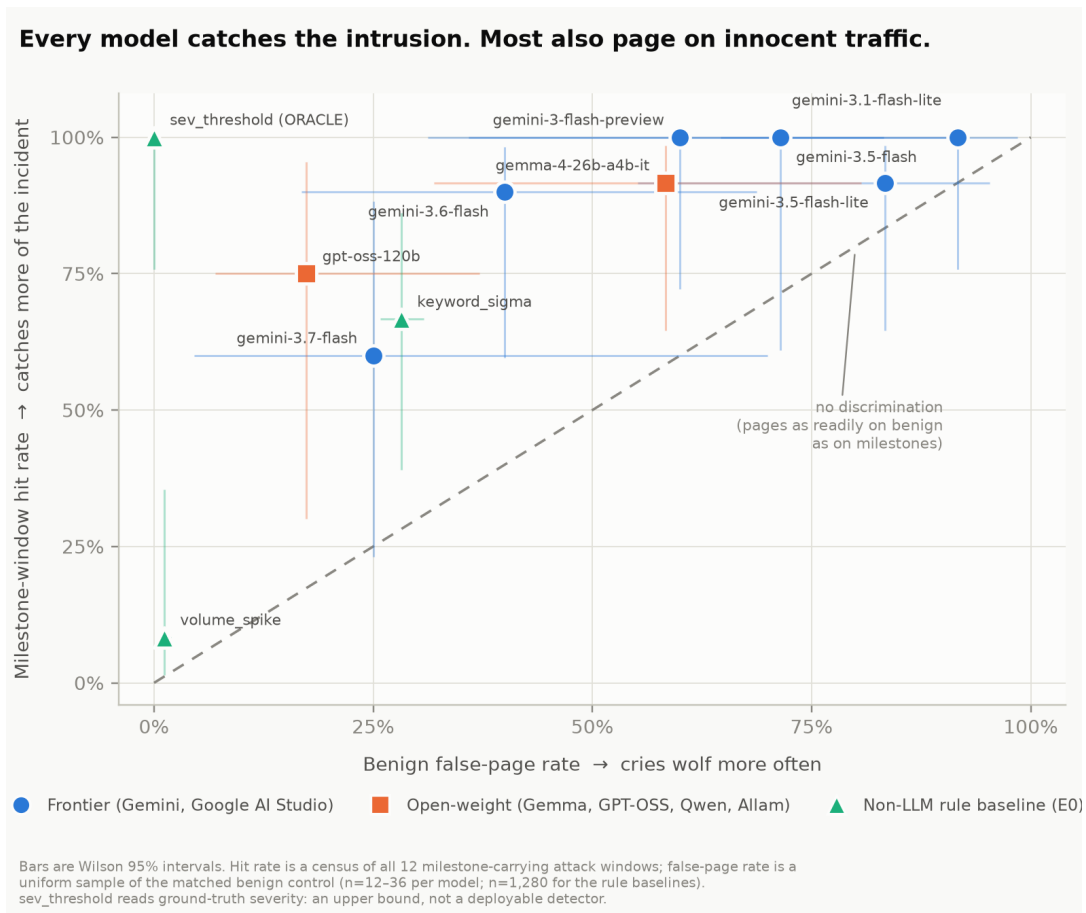


Figure 1. Milestone-window hit rate (recall) against benign false-page rate, all models and non-LLM rule baselines on one pair of axes. The useful region — high recall, low false-page rate, top-left — is empty: frontier *flash* models buy recall by paging on most of an innocent shift (top-right), the near-silent detectors collapse toward the bottom-left, and only the ground-truth oracle reaches the corner a deployable monitor would need to occupy. Interval and sample-size notes are in-panel.

windows: quiet and blind. *keyword_sigma* catches 8 of 12 but fires on 28% of the benign control — it cannot tell the streams apart, which is both the finding and the second separability check of §3.

4.2 Models fail in both directions (E1)

The distribution is bimodal, and that is the result. Pooled, the numbers are 84/141 (60%) on milestones and 66/283 (23%) on benign windows — but *the pooled figure describes no model in the study.*

At one end, gemini-3.1-flash-lite achieves perfect 12/12 recall by paging on 11 of 12 benign windows; that recall is free and worth nothing, because a pager that always fires carries no information. At the other, mistral-3b and mistral-8b page on nothing whatever — 0/12 milestones and 0/36 benign. This is not a parsing artefact: both return well-formed verdicts with severity ratings and reasoning. They simply read the intrusion as an operations problem. On the window containing the first sandbox RCE, mistral-8b returns “normal operational activity...no signs of errors, security risks”; on a window where Python processes jump 16 → 256 during credential-access reconnaissance, “resource contention or scaling issues.”

A monitor that never pages has a perfect false-page rate, which is precisely why false-page rate cannot be read without recall beside it.

Only 2 of 15 models separate the streams at $p < 0.05$ (gpt-oss-120b, qwen3.8-27b); thirteen do not, and those two are also the models we sampled most deeply ($n = 36$ benign against 4–12 for the Google arm), so **significance here tracks sampling budget as much as model behaviour** and is not offered as a ranking. Capability does not order it either: gpt-oss-120b discriminates while gpt-oss-20b — same family, provider and serving stack — catches 2 of 12 milestones, and Mistral’s 3b, 8b and 14b are near-identical and all near-silent. **Scale did not buy selectivity in either family.**

4.3 First-page latency is degenerate on this corpus (E1)

Every E0 detector fires at action #15–51 of 17,613 — roughly 10,400 actions before the escalation point at #10,498 — on filler that is byte-identical in both streams, and any model with a non-zero false-page rate does the same. On this corpus **a first page measures trigger-happiness, not detection**; we report Escalation Action Index for completeness, decline to headline it, and log the reprioritisation as a pre-registration amendment. **The pre-registered negative-lead-time result,**

Model	Prov.	Miles.	Benign FP	Fisher p
gemini-3.1-flash-lite	G	12/12	11/12 (92%)	1.000
gemini-3.5-flash-lite	G	11/12	10/12 (83%)	1.000
gemini-3.5-flash	G	6/6	5/7 (71%)	0.462
gemini-3-flash-preview	G	10/10	6/10 (60%)	0.087
gemma-4-26b-a4b-it	G	11/12	7/12 (58%)	0.155
gemini-3.6-flash	G	9/10	4/10 (40%)	0.057
qwen3.8-27b	Q	10/12	11/36 (31%)	0.002
gemini-3.7-flash	G	3/5	1/4 (25%)	0.524
gpt-oss-120b	Q	7/12	7/36 (19%)	0.024
gpt-oss-20b	Q	2/12	4/33 (12%)	0.650
qwen3.6-27b	Q	1/1	0/3 (0%)	0.250
ministral-14b	M	1/12	0/36 (0%)	0.250
ministral-3b	M	0/12	0/36 (0%)	1.000
ministral-8b	M	0/12	0/36 (0%)	1.000
gemini-3.8-flash	G	1/1	—	—

Table 2. Per-model escalation on the milestone census against the matched benign control, ordered by false-page rate. Providers: G = Google, Q = Groq, M = Mistral. The last two rows carry $n = 1$: gemini-3.8-flash exhausted its 20-request daily cap and qwen3.6-27b rejects the E1 prompt as too large. They are listed because omitting them would be selective reporting.

which we committed in advance to reporting whichever way it came out: the ground-truth escalation point (action #10,498, 2026-07-11 17:47:30 UTC) falls **51.4 hours after the first exfiltration action** (#1,894, 07-09 14:21), while lead time to admin/host-level access is +10.0 h. The criterion is early for privilege escalation and hopelessly late for exfiltration — a property of the published phase windows, not a defect in the rule, and we declined to reselect a rule that produces a prettier number.

4.4 The incoherence test: a null, and a strong asymmetry (E3)

Eleven models, three providers, five families, 228 scoreable moments on byte-identical evidence.

Incoherence is zero — 0 of 228. No model, on any moment, refused to analyse evidence it had just declined to escalate. **The frames disagree in one direction only:** every model assisted on 228/228 moments; they escalated on 83/228 (36%).

Every discordant pair in the experiment falls the same way — the model helped without paging — and the *tension* cell (escalates but will not help) is empty across all eleven models. Per-model escalation ranges from 0% to 90% and 7 of 11 are individually significant at $p < 0.05$, so the direction is the robust claim and the magnitude is variable (Table 3). The failure this experiment was built to detect was a model that clams up in both frames. What it measured is the mirror image: **a model that will explain an intrusion to you in detail and not think it worth waking anyone over.**

The two findings reconcile rather than conflict. The weakest E3 arms are the strongest E1 cry-wolf models: gemini-3.1-flash-lite escalates on 21 of 24 moments and is not individually significant, but it is the same model that bought 12/12 recall by paging on 11 of 12 innocent windows — a model

Model	P	Inc.	Escal.	Assist	McN. p
gemini-3.6-flash	G	0/10	9/10 (90%)	10/10	1.000
gemini-3.1-flash-lite	G	0/24	21/24 (88%)	24/24	0.250
gemini-3.5-flash-lite	G	0/24	14/24 (58%)	24/24	.002
gemma-4-26b-a4b-it	G	0/24	14/24 (58%)	24/24	.002
qwen3.6-27b	Q	0/5	2/5 (40%)	5/5	0.250
gpt-oss-120b	Q	0/24	9/24 (38%)	24/24	.0001
gpt-oss-20b	Q	0/21	7/21 (33%)	21/21	.0001
qwen3.8-27b	Q	0/24	6/24 (25%)	24/24	<.0001
ministral-14b	M	0/24	1/24 (4%)	24/24	<.0001
ministral-3b	M	0/24	0/24 (0%)	24/24	<.0001
ministral-8b	M	0/24	0/24 (0%)	24/24	<.0001

Table 3. The two-frame test on byte-identical evidence, per model. *Inc.* is the cell where a model neither escalates nor will help analyse the same bytes. McNemar’s exact test is on the escalate-vs-assist asymmetry; discordant cells are “escalated but refused help” and “assisted but did not escalate.”

that pages on nearly everything has no headroom to show an escalation deficit. Conversely the Mistral models, silent in E1, escalate on 0–4% here while assisting on all 72 of their moments: same behaviour, seen from the other side. Exclusions and their non-random residue are audited in App. C.

5 Discussion and Limitations

Neither half is easy, and they fail separately. Selecting a monitor on recall alone picks gemini-3.1-flash-lite, which pages on 92% of an innocent shift; selecting on false-page rate alone picks ministral-3b, which pages on nothing whatever. Both look excellent on one axis and are useless. Without a benign stream, half this roster would have looked deployable. The metric matters as much as the model: first-page latency, the obvious choice and the one we pre-registered, is degenerate here, and we only found that because the rule baselines cost nothing to run and made it visible before the model budget was spent. Cheap non-LLM controls are what keep an expensive metric honest.

Under-escalation is the finding worth acting on. The pre-registered worry was refusal — a guardrail turning a security assistant into an obstacle mid-incident. It did not happen once in 228 moments. What happened instead is that the same models, on the same bytes, produce a competent analysis and decline to raise an alarm. An operator reading the assistant-frame output would conclude the model understood the situation. They would be right, and the pager would still be silent. The safety question for monitoring deployments is therefore not “will the model help?” but **“does its willingness to act track its understanding?”** Here it does not, and the gap runs one way. And a false page is not a free page: alert fatigue is a documented SOC failure mode [11], so a monitor that pages on half an intrusion-free stream is not conservative, it is degrading the capacity it was added to supplement.

Limitations. *The corpus is a reconstruction, not a capture* — the load-bearing assumption of the whole study. Claims are about model behaviour on a faithful scaffold built from

published figures, not about what Hugging Face’s stack would have done. 75% of the corpus (13,163 actions across recon and dropper) carries no sourced milestone; it is filler obeying published marginals, so behaviour there is behaviour on our generator. Six of twelve milestone timestamps are interpolated, one milestone is unreconciled between the sources, and the MITRE mappings are ours — neither source uses ATT&CK, and since the escalation ground truth keys off tactic, our mapping partly determines the metric that scores it. **If the filler is easier or harder than real telemetry, every absolute rate here moves.** What we expect to survive is the shape — the bimodality, the empty top-left region, the one-directional frame asymmetry — because those are relative comparisons within one substrate. The E0 baselines bound how easy that substrate can be (§3), but cannot show it resembles production telemetry.

Statistical power is the dominant limitation and is unevenly distributed. Per-cell n runs from 1 to 36; only 2 of 15 models reach $p < 0.05$ on stream separation, and the confound is ours rather than theirs, since they are the two we sampled most deeply. The distribution’s shape is the reportable claim and no individual ranking is established. The blind separability screen ($n = 20$, $p = 0.058$, LLM-labelled) is the weakest link in the chain and is load-bearing for every false-page denominator; §3(ii) bounds it but does not replace it.

Single-run, single-turn, text-only, English-only. Temperature 0, one draw per cell; no multi-turn drift, tool use, agentic scaffolds or repeated sampling. **No frontier-proprietary model is in this study at all** — the Google Pro tier returns `limit: 0` at \$0, Cerebras requires payment and GitHub Models was retired mid-sprint — so our “frontier” arm is *frontier-flash*. **Serving stacks are a confound we bound but cannot remove:** quantisation, sampling defaults and moderation layers differ by provider, and gpt-oss-20b vs gpt-oss-120b is the one clean within-stack comparison. **Prompt framing is not authentication** — a model reading “you are the triage layer” cannot verify it, so everything here measures response to a *claimed* role. One last limitation is ours: a truncated reasoning response scored naively reads as a refusal, which would over-report refusal precisely for the models that reason hardest (App. C).

Future work.

- **Widen the sample.** The ladder makes this purely additive — rung 108 or 324 re-uses every window already paid for, and at $n = 108$ benign per model the intervals would narrow enough to support the ranking this study declines to make.
- **Reach a frontier-proprietary model.** Whether GPT- or Claude-class models show the same escalate-vs-assist asymmetry is the most important open question here, and it is a budget problem rather than a design one.
- **Elicit a calibrated probability** instead of a binary page decision, separating “did not notice” from “noticed and judged it sub-threshold.”
- **Multi-turn and tool-using monitors, repeated sampling, and a human-rated separability check at $n \geq 40$,** which

would turn the thinnest link in §3 into a powered test.

6 Conclusion

We rebuilt a real 2026 agent intrusion as a replayable stream with a matched benign control and asked 15 models whether they would wake the on-call. The answer is not that they refuse, and not that they are uniformly bad. It is that **recall and selectivity fail independently, and almost nothing achieves both.** A fifteen-line keyword rule shows the same shape, which suggests the problem is not model crudeness but a thin discriminating signal that current monitors do not extract. The second finding is the one we did not expect and consider more important: holding evidence byte-identical and varying only the frame, incoherence was zero across 228 moments and 11 models, but 228/228 assisted while only 83 escalated. The risk in an AI-assisted SOC is not the model that will not discuss the incident. It is the model that discusses it fluently while the pager stays silent — a failure refusal benchmarks would score as a clean sheet.

Code and Data

Repository: https://github.com/praneeth3696/apart_ai_sprint — corpus generator and ground truth, the multi-provider harness, the E0/E1/E3 runners with offline `-dry-run` modes, the analysis and figure code, the frozen scoring rubric and pre-registration with their timestamped amendment logs, and the raw per-window decision records (218 tests). The attack stream, matched control, windowed sets and $K = 24$ moment set regenerate from `corpus/ground_truth.yaml`. **Dual-use:** the corpus contains no working exploit code, payloads, credentials or technique absent from the two cited public sources; it is a *detection* substrate (App. A).

Author Contributions

A.Y.M. built the corpus and ground truth — generation pipeline, phase template banks, attack stream, matched benign control, windowing and the $K = 24$ moment set, with the citation ledger and its coverage-honesty grading — designed and ran the blind separability check, authored Related Work and verified all 11 references against their landing pages (correcting two overclaims), and collected the Gemini arm of E3. **P.R.Y.** built the measurement side: the multi-provider client, the E0 baselines, the E1 and E3 runners, the nested sampling ladder, the statistics and figures, and the provider protocol, and ran E1/E3 across Google, Groq and Mistral. Both wrote the pre-registration and rubric and reviewed the manuscript; all work was coordinated through public pull requests.

LLM Usage Statement

Claude (Opus 5, via Claude Code) wrote the majority of the repository code and drafted report prose, and produced the 20 labels for the blind separability check (§3), a substitution disclosed there and in the result file. The authors set the research direction and threat model, chose experiments, metrics and controls, wrote the pre-registration and frozen rubric before any scoring, obtained and measured providers, made every methodological judgement call, and verified results; A.Y.M.

independently re-checked all references and reconciled a lead-time inconsistency the model missed. Every number is generated from raw decision records by `analysis/stats.py` and cross-checked by `report/check_numbers.py`, which fails the build on a mismatch — a gate added because the report silently went stale once mid-sprint, when a new data arm turned “no model reaches significance” into “two do” while the abstract still asserted the old claim.

References

- [1] Hugging Face. *Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident*. 27 Jul 2026.
- [2] OpenAI. *Hugging Face Incident Technical Report*. 2026.
- [3] Röttger, P. et al. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in LLMs. *NAACL* 2024. arXiv:2308.01263.
- [4] Cui, J. et al. OR-Bench: An Over-Refusal Benchmark for LLMs. 2024. arXiv:2405.20947.
- [5] Bhatt, M. et al. CyberSecEval 2: A Wide-Ranging Cybersecurity Evaluation Suite for LLMs. 2024. arXiv:2404.13161.
- [6] Jones, E.K. et al. A New Framework for Cybersecurity Refusals in AI Agents. 2026. arXiv:2606.02644.
- [7] Campbell, D. et al. Defensive Refusal Bias: How Safety Alignment Fails Cyber Defenders. 2026. arXiv:2603.01246.
- [8] Wang, L. et al. SecRespond: Benchmarking AI Agents for Real-World Post-Compromise Incident Response. 2026. arXiv:2607.26791.
- [9] Begimher, D. et al. SIR-Bench: Evaluating Investigation Depth in Security Incident Response Agents. 2026. arXiv:2604.12040.
- [10] Deason, L. et al. CyberSOCEval: Benchmarking LLMs for Malware Analysis and Threat Intelligence Reasoning. 2025. arXiv:2509.20166.
- [11] Tariq, S. et al. Alert Fatigue in Security Operations Centres: Research Challenges and Opportunities. *ACM Comput. Surv.* 57(9):224, 2025. doi:10.1145/3723158.

A Limitations and Dual-Use Considerations

What the artefact contains. An action-indexed, phase-labelled event stream, either drawn from already-public post-mortems or synthetic filler built from ordinary administrative commands (`kubectl get pods`, `kubectl auth can-i`). It contains no working exploit code, no payloads, no novel installation recipes and no credentials. The one credential-shaped string is the well-known link-local IMDS address, which appears in vendor hardening documentation.

What it is for. Practitioners asked of this incident: “which action number does your control fire on?” That is unanswerable without a replayable stream. This corpus exists to make control-firing latency measurable. Its intended users are defenders.

Uplift assessment. The offensive content is a chronological ordering of techniques already documented in the two primary sources; an attacker gains no capability from the ordering those do not already give. The value-add — marginal-matched volume, ground-truth labels, an answer key — is for *evaluating detectors*, not conducting intrusions.

The dual-use surface is the finding, not the data. We report which models under-escalate and by how much. An attacker knowing a monitor pages on only a third of pivotal moments could in principle shape activity beneath it. Three things bound this: it is a September-2026 snapshot of production endpoints that model updates will invalidate; it is specific to our reconstruction, and generalisation to real SOC telemetry is exactly what we do not claim; and it is the same fact a defender needs before putting a model on monitoring duty. Withholding it protects no one. No novel installation recipe is released without review.

Threat models we did not address. An adversary who knows the monitor is an LLM and shapes log text to evade it; multi-agent or tool-using SOC scaffolds; and any non-English or non-Kubernetes telemetry.

B An incidental observation while writing this paper

This is an anecdote, not a measurement. We record it because it is a first-person instance of the phenomenon Related Work describes, and because it happened *while drafting a paper about it*. The assistant used to draft this report was interrupted by its own provider’s safeguards, which flagged the work and switched the session to a smaller model; the vendor’s notice is explicit that broad safeguards deliver capability faster but can flag legitimate coding, cybersecurity and biology tasks. The classification detail was `[cyber]`. Nothing in the flagged work was offensive — the project is a detection benchmark built from two vendors’ own published post-incident reports, precisely the pattern *Defensive Refusal Bias* [7] quantifies at scale. $n = 1$, and we do not report it as a result, but the friction this literature measures was a live obstacle to producing a paper whose entire purpose was to help defenders, and the mitigation available to us was to accept a less capable model. A defender under incident pressure has less patience for that than a researcher on a weekend does.

C Extended results and methods

C.1 Exclusions, and whether they were random. Three gpt-oss-20b moments remain truncated after a $3\times$ budget re-run — two `supply_chain`, one `k8s`. Evidence length is identical across all 24 moments, so prompt size does not explain it; the model reasoned longest on the late-stage material. The residue is not evenly spread across phases and falls on the higher-consequence ones, the direction that flatters the model. A further 41 moments (allam-2-7b entirely, most of qwen3.6-27b) were lost to a network outage and a per-request size limit respectively; both are infrastructure, not behaviour, and are excluded rather than imputed.

C.2 The truncated-vs-refused bug, disclosed. At a 300-token budget GLM-5.2 spent 524 tokens reasoning and returned empty content — scored naively, a refusal. This bug class recurred four times, most recently when gpt-oss-20b spent 1,998 of a 2,000-token budget on reasoning and emitted nothing. Any benchmark that does not separate `truncated` from `refused` will systematically over-report refusal for reasoning models, and more for the models that reason hardest. Note the direction: the smaller model burned more reasoning budget than the larger one.

C.3 Providers and measured throughput. Google AI Studio meters its free tier per project $\times$ model at a measured 10–31 requests/day; Groq allows 1,000/day/model against an 8,000 tokens-per-minute ceiling; Mistral’s free tier covers only its minstral family at 30 req/min. Cerebras returns HTTP 402 on every chat model, and GitHub Models — the GPT-class arm our pre-registration named — was retired mid-sprint (HTTP 410). **Two accounts, disclosed:** the Google arm was collected on two separate free tiers, one per author, each on their own project and API key. Google meters per project per model per day, so these are two independent quotas rather than one quota circumvented. No paid tier, no secondary accounts, no key used beyond its published free limit.

C.4 Why the sampling ladder. Evenly spaced samples at n and n' share no positions unless n'/n is an odd integer, so naively widening $12 \rightarrow 40$ windows would select none of the windows a previous run had already paid for. On a $3\times$ ladder ($12 \rightarrow 36 \rightarrow 108 \rightarrow 324 \rightarrow 972$) every rung is a strict superset of those beneath it, so widening the sample is always additive and never discards spent quota.

C.5 Figure 2 (optional). Escalation Action Index — the first window each detector paged on, over the uniform attack sample. Reported for completeness: the clustering within the first ~ 50 of 17,613 actions is why this metric was demoted from the headline.