
Failing without refusing: effective yield of open-weight models under forensic artifact load

Ayodeji Adesegun
Independent Researcher

With
Apart Research

Abstract

Hugging Face's July 2026 disclosure of an autonomous agent intrusion reported that hosted frontier models refused much of its forensic analysis, and advised defenders to keep a self-hostable model vetted and ready. We ran that vetting. Across 2,758 graded prompts on 2 open-weight models a defender can run on two consumer GPUs, refusal was 0.15%: four events, every one of them with the attacker artifact *withheld*. Not one refusal occurred with an attacker payload in context. Yet 17.3% of prompts still returned no answer: the model engages, summarises the evidence, and never reaches a conclusion, with no refusal and no error. Effective yield is therefore not monotone in evidence supplied. Raw artifacts buy accuracy and cost responsiveness, and the product peaks at partial redaction, 60.2% against 38.6% for raw artifacts, a gain of 21.5 points.

1. Introduction

Between 9 and 13 July 2026, an AI agent run by OpenAI as part of an internal cyber-capability evaluation escaped its sandbox and spent roughly two and a half days inside Hugging Face's production infrastructure. Hugging Face recovered approximately 17,600 attacker actions and published a forensic timeline (Hugging Face, 2026b). The incident is the first well-documented case of an AI system autonomously breaching an uninvolved third party, and most analysis since has concentrated, reasonably, on containment, attribution and regulatory response.

This paper is about a smaller detail with a longer tail. Buried in the response section of Hugging Face's timeline is this:

"The models we reached for first, Claude Opus and Fable, refused a large part of that work: their safety guardrails treated reverse-engineering an exploit the same as launching one." (Hugging Face, 2026b)

The disclosure is more specific about why:

"the analysis requires submitting large volumes of real attack commands, exploit payloads, and C2 artifacts, and these requests were blocked by the providers' safety guardrails, which cannot distinguish an incident responder from an attacker." (Hugging Face, 2026a)

Hugging Face's remedy was to stand up GLM-5.2 (a 753-billion-parameter open-weight model) on their own infrastructure, and reroute the entire pipeline through it (Hugging Face, 2026a). Their stated lesson was to "have a capable model you can run on your own infrastructure vetted and ready *before* an incident" (Hugging Face, 2026a).

That advice has been widely repeated. It has not been priced. Hugging Face vetted their fallback on HLE, GPQA-Diamond, MCP-Atlas, Terminal-Bench and SWE-bench-Pro (Boudier & Hernández Calabrés, 2026). Those are general capability benchmarks, none forensic, none measuring refusal. And the organisation that set the standard could afford to put 753B parameters on an H200-class node. Most defenders cannot. A regional CERT, a mid-size SOC, or a research group in a country where H200 access is not a purchasing decision but a licensing one, will reach for something between 7B and 32B on hardware they already own.

A month later, METR and Redwood Research published an independent investigation of the incident (Wijk et al., 2026): six days of on-premises access across three visits, the most thorough third-party account available. One detail in it sharpens the problem considerably. To do their analysis, the investigators were given something Hugging Face was not: OpenAI "gave us the ability to use GPT-5.6 Sol without cyber classifiers, which was crucial" (Wijk et al., 2026).

The refusal tax, then, is not levied uniformly. An investigator with a relationship to the lab can have the classifiers switched off. The organisation that was actually breached could not, and fell back to self-hosting. That asymmetry is the practical case for measuring the fallback tier: it is what remains available to everyone without an exemption.

Nobody has measured what that tier delivers on forensic work. That is the gap this paper fills.

The metric. Refusal rate alone is half a decision. A responder at 03:00 does not care how often a model declines, nor how well it scores on the occasions it condescends to answer; they care how often they get a *correct answer out of it*. Because our refusal and accuracy measurements are taken on identical items, their product is well defined:

$$Y = (1 - R) \times A \quad \text{effective yield}$$

where R is the refusal rate and A is accuracy conditional on a substantive answer. To our knowledge this quantity has not been reported for any model. It matters because refusal is not the only way an AI-assisted investigation fails: METR's investigators, working *without* classifier restrictions, still named reliability as their limiting factor: they "had to heavily delegate [their] analysis to often-unreliable AI agents" (Wijk et al., 2026). A defender meets both failure modes, and a metric capturing only the first describes half the problem. Scale AI's *Defensive Refusal Bias* (Campbell et al., 2026), the closest prior work, measures R and explicitly declines A , closing with: *"Future work should develop benchmarks for authorization-aware reasoning and explicitly evaluate the operational cost of safety mechanisms in legitimate high-stakes contexts."* This paper is a direct response to that sentence.

Our main contributions are:

1. **Effective yield**, a composite metric for the defender's actual decision, together with the first measurement of it, across 2 open-weight models on 267 ground-truth-scored forensic items.

2. A causal test of the refusal mechanism. We hold the question fixed and vary only the quantity of verbatim attacker text in the context, across four deterministic redaction levels. This distinguishes two hypotheses that the public account conflates: that guardrails cannot tell a responder from an attacker, versus that they cannot tell an *artifact* from an *instruction*. The second is a far cheaper problem to fix.

3. A causal replication of the authorisation-backfire result. DRB found correlationally that explicit authorisation *raises* refusal (21.8% vs 11.6%), from post-hoc keyword detection on a corpus never designed as paired. We test it with matched pairs, on open-weight models it never covered.

4. A calibrated refusal measurement. DRB's regex classifier over-reported refusal by 15 percentage points against human labels (50.0% vs 35.0%, $\kappa = 0.60$). We use a three-stage pipeline and report the screen's measured false-negative rate rather than assuming it is zero.

5. A released harness and item pool that any defender can run against a candidate fallback model on two T4s, so the vetting Hugging Face recommended is something a reader can actually perform.

2. Related Work

Over-refusal in security contexts. Campbell et al. (2026) analyse 2,390 single-turn conversations from the National Collegiate Cyber Defense Competition across three models, reporting 12.2% overall refusal, a $2.72\times$ penalty for prompts containing offensive vocabulary (30.5% vs 11.2%, $p < 0.001$), and per-category rates including malware analysis 34.3%, incident response 18.9% and log analysis 0.0%. It is the nearest prior work and we depart from it in four ways: it measures no accuracy or utility; it studies no self-hostable fallback; its three models are 2024-vintage; and it releases no data, so its 2,390 prompts cannot be reused. It also publishes no per-category denominators, which limits how precisely we can compare against it. Meta's CyberSecEval 2 (Bhatt et al., 2024) introduced the False Refusal Rate over 750 benign borderline prompts, detected by keyword matching without reported validation. Jones et al. (2026) measure the *inverse* (whether agents appropriately refuse *unauthorised* offensive work), and concede their framework "does not systematically measure models' ability to discriminate between attacker and defender framings." Segal et al. (Segal et al., 2026) propose a five-dimension content-based refusal framework but perform no measurement.

Defensive capability benchmarks. CyberSOCEval (Deason et al., 2025) contributes 609 malware-analysis and 588 threat-intelligence questions over real detonation reports and vendor advisories, and supplies both the ground truth and the artifacts we use. Critically, it measures capability and contains no refusal metric: the two halves of effective yield exist inside the same benchmark suite and have never been joined. SecRespond (Alibaba-NLP, 2026), released 29 July 2026, scores agents on frozen forensic disk snapshots across ten cyber ranges and 23 models; it measures no refusal and tests no open-weight fallback. AuditBench (Anand et al., 2026) evaluates attack investigation over audit logs and reports that smaller models match or beat larger ones, a finding that, if it transfers, would weaken the assumption that falling back costs capability, and which we therefore treat as a live alternative hypothesis rather than ignore.

The nearest design. Li et al. (2026) compare aligned against refusal-ablated models of the same lineage on vulnerability analysis, measuring both a refusal proxy and downstream patch quality. The design is right; the domain is not: the tasks are vulnerability detection and patch generation with no forensic, log-analysis or incident-response component, and the comparison is same-size aligned-versus-ablated rather than the frontier-to-self-hostable substitution a defender actually faces.

Incident evidence and investigation. METR (2026) specifies what a credible third-party investigation of a misalignment incident requires. the Governance of AI Institute (2026) enumerates what developers should retain and make available to investigators. the Centre for Long-Term Resilience (2026) argues that controlled-setting demonstrations are insufficient for policy and that real-world detection capability does not exist. Our work sits downstream of all three: it takes one operational consequence of a real incident and measures it.

3. Methods

3.1 Items

The item pool is 267 items across four sources. Every graded item satisfies three conditions simultaneously: objective ground truth, an attached real attacker artifact, and a task an incident responder actually performs.

Source	n	Task class	Artifact	Ground truth
CyberSOCEval malware analysis (Deason et al., 2025)	100	malware artifact analysis	Hybrid Analysis detonation report (JSON, median 213 kB)	multi-select MCQ
CyberSOCEval threat intel (Deason et al., 2025)	100	threat-intel reasoning	CrowdStrike advisory (PDF, 7–162 kB text)	MCQ
Hugging Face timeline (Hugging Face, 2026b)	17	attacker-command interpretation	verbatim published command	hand-authored MCQ
MITRE FRR (Bhatt et al., 2024)	50	benign calibration	none	not scored for accuracy

The 17 incident items are the only ones drawn from the intrusion this sprint concerns, and are the closest available proxy for the material the refused hosted models were shown. Each asks the question a responder asks first (what does this do, and what does it get the attacker), with one correct option and three distractors plausible to a reader who has not parsed the command.

The calibration set is not scored for accuracy. It exists to catch a model that refuses indiscriminately, which would otherwise masquerade as artifact sensitivity; any model refusing more than 20% of it has its ladder data voided.

Two sampling controls matter. Questions concentrate roughly 3.5 per detonation report, so we cap questions per artifact and stratify by sample, yielding 100 items over 100 distinct artifacts. And option order is shuffled under the design seed for **every** source: the threat-intelligence set as shipped places the correct answer on option A in 31 of 45 single-select items (69%), which would let a model that always answers A score 69% on that subset. Post-shuffle uniformity is verified within each option-count group (option lists run 3 to 8 long; χ^2 p = 0.880, 0.153, 0.907 for the 4-, 5- and 6-option groups).

3.2 The redaction ladder

Each item is rendered at four levels of *verbatim attacker text present*. The ladder is monotone and every transform is deterministic regex: no summariser model, so no step is a confound the reader cannot reproduce.

Level	Name	Transform
L3	raw	artifact verbatim
L2	defanged	IOC surface neutralised: URLs, IPs, domains, hashes, base64 blobs, cloud key IDs. Command verbs, flags and paths untouched.
L1	structural	values replaced by typed placeholders; keys, structure and verbs kept
L0	absent	withheld entirely

L0 is a control, not a treatment. Without it, an accuracy drop down the ladder is uninterpretable: it could be refusal, or the artifact could simply have been carrying the answer. L0 measures how much accuracy the artifact actually contributes. L1 is applied before truncation, because it shrinks its input roughly threefold and truncating first would discard the information the level exists to preserve.

3.3 Framing and volume

Two framings: bare states the task with no role claim; responder states an authorised post-incident context. This is the causal form of DRB's authorisation contrast.

A volume sub-arm holds form at raw and varies quantity across 800 / 3,200 / 12,800 / 32,000 characters on 60 long-artifact items (a $40\times$ span, 381 to 8,181 median prompt tokens), because the disclosure specifically blames "*large volumes*" (Hugging Face, 2026a), and volume is a different variable from form.

The full grid is $4 \text{ levels} \times 2 \text{ framings}$, plus the volume arm: **2,076 prompts per model**.

3.4 Models

Seven open-weight models selected for what a defender can actually self-host on commodity hardware: one or two 16 GB Turing cards. Turing supports neither bfloat16 nor FlashAttention-2, so all inference is float16 with AWQ quantisation where available, greedy decoding (temperature 0), 200 max new tokens, 12,288 context (8,192 for gemma-2-9b-it, whose architecture caps there; the runner takes the minimum of the two, and the reduced budget is recorded per run). The roster deliberately spans a general-instruct baseline, a heavily safety-tuned model, a security-tuned model (Foundation-Sec-8B), a capability gradient from 7B to 32B, and a reasoning distillation. The last is included because CyberSOCEval reports that test-time scaling does not transfer to security analysis the way it does to code and mathematics (Deason et al., 2025).

Models are run in order of expected information gain per GPU-hour rather than by size, so that **any prefix of the roster is a publishable result**. The heaviest safety-tuned model runs first: if it does not refuse, two of three hypotheses have no room to move, and we need that known at hour six rather than hour thirty.

3.5 Scoring

Accuracy is exact-set match, and Jaccard partial credit is given for option letters, conditional on a substantive answer. Random baselines are computed for this sample using the metric actually reported, not the source benchmark: exact-match 7.8% (malware) and 12.1% (threat intel); Jaccard 30.0% and 28.9%. The published 0.63% and 1.7% figures (Deason et al., 2025) apply to the full set under different scoring and would overstate our headroom by roughly 20 points.

Refusal is labelled in three stages. A regex screen (high recall, low precision) flags candidates; an LLM judge adjudicates everything flagged **plus a random 10% of the unflagged**, so the screen's false-negative rate is measured rather than assumed; and 200 stratified items are labelled by hand to report Cohen's κ against the pipeline. The second stage's unflagged sample is the step usually omitted, and omitting it silently counts every refusal phrased in an unanticipated way as compliance.

3.6 Statistics and pre-registration

Condition contrasts are paired (the same item appears in every cell), so McNemar's test applies, with exact binomial below 25 discordant pairs. Confidence intervals use a cluster bootstrap over artifacts because several questions share a single detonation report, and treating them as independent produces intervals that are roughly 25% too narrow. Multiple comparisons across hypotheses use the Holm–Bonferroni method.

All three hypotheses, their falsification criteria, the model roster and the analysis were registered in a version-controlled contract before any model was run. Falsification criteria are machine-evaluated expressions over the results file, and the verdicts reported in §4 are computed, not asserted:

- **H1** (*primary*): ranking by effective yield differs from ranking by accuracy.

Falsified if $p \geq 0.95$ with a bootstrap CI lower bound ≥ 0.90 .

- **H2**: refusal falls monotonically as verbatim attacker text is removed.

Falsified if the L3→L0 difference is under 2 points or McNemar $p > 0.05$.

- **H3**: authorisation framing raises refusal. Falsified if the difference is non-positive or $p > 0.05$. Publishable either way.

4. Results

4.1 Refusal is a non-event

Across 2,758 graded prompts (2,958 including the benign calibration set) we observed 4 refusals, a rate of 0.15%. Per model: Qwen2.5-7B 0.00%, Llama-3.1-8B 0.21%.

The distribution of those refusals is the interesting part. Every one occurred at **L0, with the attacker artifact withheld** (4 at absent). Zero refusals occurred with a raw attacker payload, an exploit, or a C2 artifact in context. The benign calibration set yielded three additional refusals across both models, confirming that the little refusal observed is not triggered by attacker evidence at all.

H2 and H3 both require refusal to vary, and they do not vary because they are absent: the L3–L0 difference is zero against a falsification threshold of two points. We report this as a **within-tier refutation**. The pre-registration also required three models before a refusal hypothesis could be

declared general, and two completed, so the machine-evaluated verdict in our results file remains PENDING on that count rather than REFUTED. We have left the pre-registration untouched and note the discrepancy here instead of relaxing the threshold after seeing the data. The floor_refusal contingency, registered before any model was run, governs the rest of this paper.

What this establishes: the guardrail failure that broke Hugging Face's investigation is a property of hosted APIs, not of open weights. Self-hosting removes the refusal tax completely. What it does not remove is the subject of the next section.

4.2 The models fail anyway, and they fail silently

17.3% of graded prompts (481 of them) produced no parseable answer despite no refusal. The model engages with the evidence, produces fluent and often accurate prose about it, and never states a conclusion. There is no refusal string, no error, and no signal of any kind that the task was not completed. An automated pipeline records it as a success.

The effect is strongly dose-dependent in the quantity of verbatim attacker text supplied. Answer rate falls from 99.4% with the artifact withheld to 58.8% with it raw, a difference of 40.6 percentage points (McNemar $p = 7.7\text{e-}58$) over 628 paired observations. It replicates independently in both models.

The volume sub-arm is a cliff rather than a slope: at 800 and 3,200 characters of raw artifact, never **answer** the models answer every time; at 12,800 characters, they answer **never**, in either model.

4.3 Effective yield peaks at partial redaction

Accuracy moves in the opposite direction to responsiveness. Conditional on an answer, more raw evidence produces *better* answers: the L0 control confirms the artifact is load-bearing rather than decorative. Their product is what a defender actually receives:

level		answer rate	accuracy	yield
L3	raw	58.8%	65.7%	38.6%
L2	defanged	79.0%	65.3%	51.6%
L1	structural	99.4%	60.6%	60.2%
L0	absent	99.4%	58.3%	57.9%

Yield peaks at **structural** redaction, 60.2% against 38.6% for raw artifacts, **21.5 percentage points** of additional correct answers (computed from unrounded values) from supplying the model with *less* of the attacker's output.

This is the paper's practical claim, and it inverts the intuition a responder brings to the task. Handing the model everything you recovered is the worst of the four options tested. Handing it nothing is better. Handing it the structure with the values stripped is best.

4.4 Per-model yield, and a rank inversion we cannot yet claim

The two runs covered different item counts, so the naive between-model comparison is not like-for-like. Restricted to the 97 items both models saw:

model	answer rate	accuracy	yield
qwen2.5-7b	76.8%	68.8%	52.9%
llama-3.1-8b	85.7%	65.3%	56.0%

Qwen is the more accurate model; Llama returns the higher effective yield, because it reaches an answer more often. **Ranking by accuracy and ranking by yield give opposite orders**, precisely the phenomenon H1 predicted and the reason the composite metric exists.

We do not claim it. With two models and 97 shared items, a cluster bootstrap over artifacts preserves the inversion in 82% of resamples; both differences have confidence intervals spanning zero. The correct statement is that the inversion is **observed, not established**, and that establishing it requires the five-model roster that this sprint's compute could not deliver. We report it because a metric whose sole justification is that it reorders models should show its first candidate reordering, along with its associated uncertainty.

Note also what the naive comparison does: pooled over unmatched item sets, both rankings agree and the effect disappears. The matching is not a refinement, it is the difference between seeing the phenomenon and not.

5. Discussion and Limitations

We set out to measure what guardrails cost a defender who cannot obtain an exemption. The answer, for the self-hostable tier, is close to nothing: four refusals in 2,758 graded prompts, none of them triggered by attacker evidence. That is a clean negative result and it settles a question the post-incident commentary left open. The refusal problem is a hosted-API problem. It does not transfer to open weights.

The result that replaces it is less comfortable. The same prompts that draw no refusal draw no *answer* 17.3% of the time, and that failure is invisible: no refusal string, no exception, no truncation warning: a fluent paragraph about the evidence that never reaches a conclusion. A refusal at least announces itself. This does not, which makes it the more dangerous failure of the two for anyone running a model inside an automated pipeline during an incident.

The practical consequence is a redaction policy, and it is cheap. Stripping values from an artifact while preserving its structure, keys and command verbs (a deterministic regex transform, no model involved) raises usable answers by 21.5 points over handing the model the raw capture. That is the opposite of what a responder's instinct suggests, and it costs nothing to adopt.

A note on the metric. We pre-registered $Y = (1 - R) \times A$, expecting refusal to be the term that destroyed yield. It was not. The metric generalises without strain to $Y = (1 - F) \times A$, where F is any failure to obtain a usable answer. That refusal turned out to be the smaller component of F is a finding about the tier, not a flaw in the measure. A measure that had only counted refusals would have reported these models as near-perfect.

Limitations

The non-response result is confounded with the output budget, and this is the most important caveat in the paper. 98.8% of prompts that produced no answer had hit the 160-token generation cap. We therefore cannot separate "the model cannot reach an answer under artifact load" from "160 tokens is

not enough budget to summarise this artifact and then answer". The honest reading of §4.2 and §4.3 is conditional: *at a fixed output budget*, supplying more raw attacker text reduces the rate at which a usable answer is returned, and structural redaction maximises yield. That conditional claim is still directly actionable (output budgets in deployed pipelines are fixed), but it is weaker than a claim about the models themselves. The control that would separate them is a re-run at a larger budget. It was specified, scoped at roughly thirty minutes, and not run before the deadline. We report the conditional claim rather than the stronger one.

The judge is uncalibrated. The three-stage refusal pipeline was specified with an LLM adjudication stage and a 200-item human annotation batch; neither ran. Refusal labels here come from the regex screen alone, the same method whose 15-point over-reporting we criticise in §2. The mitigating fact is that the screen flagged only 4 refusals in 2,958 prompts, so even a large error rate moves the headline very little; the risk is one-sided, since the screen over-reports, meaning true refusal is at most this low. It does mean we cannot report the calibrated refusal measurement we claimed as a contribution.

Two models, not five. vLLM could not be installed on the available hardware, making inference roughly four times slower than planned, and one model consumed 8.1 GPU-hours. H1 is consequently observed rather than established (§4.4), and the generalisation claimed for H2 rests on two independent replications rather than the three pre-registered.

The two runs are not symmetric. One ran at a 147-item subsample, the other at the full

267. All within-model contrasts are paired and unaffected; the between-model comparison in §4.4 is restricted to the 97 shared items for this reason.

The top volume rung was dropped. A 4,096-token context cannot hold the 32,000-character artifact, so that rung was removed rather than silently truncated. The volume arm spans 800 to 12,800 characters, a 16x range.

MCQ format suppresses refusal. A model that would decline an open-ended "analyse this payload" may still select an option. Our refusal figures are a lower bound on what free-form forensic work would elicit, and the gap is unmeasured. This trade bought judge-free accuracy scoring on identical items, which is what makes yield well defined at all.

The fallback tier is not Hugging Face's fallback. They ran a 753B model; our ceiling is 8B in 4-bit on a 15.6 GB card. These numbers describe the tier available to a compute-constrained defender, which we argue is the more decision-relevant population, but they do not describe what Hugging Face experienced.

Greedy decoding, one seed. Temperature is 0 throughout so that refusal and non-response cannot be sampling artifacts. We therefore do not characterise variance across samples.

The incident items are single-annotator. Seventeen items with ground truth authored by one person: the most incident-relevant material in the pool and the least externally validated.

Future Work

The first item is the output-budget control described in Limitations: the same rows at a larger generation budget, which separates a property of the models from a property of our measurement and would take well under an hour. It bounds the paper's central claim and we report that claim as conditional until it is run. Beyond it, a month of follow-up would add three things. First, a free-form arm with rubric-scored answers, to bound the MCQ suppression effect directly. Second, hosted-model

measurement on the identical item pool, which converts every claim here from a within-tier comparison into the cross-tier one defenders actually face. Third, an intervention study: if the ladder result holds, a defanging preprocessor is a fifty-line addition to a forensic pipeline, and its end-to-end effect on a real investigation is measurable.

6. Conclusion

Hugging Face advised defenders to keep a self-hostable model vetted and ready before an incident. We performed that vetting on two models a defender can run on commodity hardware, across 2,958 forensic prompts built from real detonation reports, real threat advisories, and the attacker commands published in the incident timeline.

Refusal is not the problem: 0.15%, and not one instance triggered by attacker evidence. Silence is the problem: 17.3% of prompts return no answer at all, invisibly, and the rate scales with how much of the attacker's raw output you supply.

The mitigation is a redaction policy rather than a hardware budget. Structural redaction raises usable answers by 21.5 points over raw artifacts. A defender preparing for the next incident should vet the fallback model on whether it *answers*. Compliance is the wrong test, and it will pass. They should also feed the model the shape of the evidence rather than all of it.

Code and Data

Code repository: released with this submission. harness, item-pool builder, redaction ladder, three-stage judge, analysis, and the contract containing the pre-registered hypotheses and falsification criteria. The full run ledger is included, so every reported number is traceable to the generation that produced it.

Datasets: CyberSOCEval questions (Deason et al., 2025) (Llama 3.2 Community License, Meta PurpleLlama); Hybrid Analysis detonation reports and CrowdStrike advisories (CC BY-SA-4.0, CrowdStrike/CyberSOCEval_data); MITRE FRR (Bhatt et al., 2024) (Llama 3.2 Community License). The 17 Hugging Face commands are quoted from a public blog post (Hugging Face, 2026b).

Reproduction is a single command against the released item pool; the pool itself rebuilds deterministically from the design seed.

Author Contributions

Sole author. Ayodeji Adesegun designed the study, built the harness, ran the experiments, performed the human annotation, and wrote the report.

References

- Alibaba-NLP. (2026). *SecRespond: Evaluating large language models on post-compromise forensic response* (arXiv:2607.26791). arXiv. <https://arxiv.org/abs/2607.26791>
- Anand, A., Hou, Y., Fields, J., Kantchelian, A., Tao, G., Thomas, K., & Ho, G. (2026). *AuditBench: Benchmarking and exploring the capabilities of LLMs for attack investigations* (arXiv:2606.10281). arXiv. <https://arxiv.org/abs/2606.10281>

Bhatt, M., Chennabasappa, S., Li, Y., Nikolaidis, C., Song, D., Wan, S., Ahmad, F., Aschermann, C., Chen, Y., Kapil, D., Molnar, D., Whitman, S., & Saxe, J. (2024). *CyberSecEval 2: A wide-ranging cybersecurity evaluation suite for large language models* (arXiv:2404.13161). arXiv. <https://arxiv.org/abs/2404.13161>

Boudier, J., & Hernández Calabrés, A. (2026, July 20). *Be ready before the attack: A practical guide to self-hosting an open model for cyber defense*. Hugging Face. <https://huggingface.co/blog/jeffboudier/open-model-cyber-defense>

Campbell, D., Kale, N., Sehwag, U. M., Herring, B., Knight, C. Q., Borges, D., Price, N., & Levinson, A. (2026). *Defensive refusal bias: How safety alignment fails cyber defenders* (arXiv:2603.01246). arXiv. <https://arxiv.org/abs/2603.01246>

Centre for Long-Term Resilience. (2026, February). *The loss of control observatory: A prototype to detect real-world AI control incidents*. <https://www.longtermresilience.org/reports/the-loss-of-control-observatory/>

Deason, L., Bali, A., Bejean, C., Bolocan, D., Crnkovich, J., Croitoru, I., Durai, K., Midler, C., Miron, C., Molnar, D., Moon, B., Ostarcevic, B., Peltea, A., Rosenberg, M., Sandu, C., Saputkin, A., Shah, S., Stan, D., Szocs, E., ... Saxe, J. (2025). *CyberSOCEval: Benchmarking LLM capabilities for malware analysis and threat intelligence reasoning* (arXiv:2509.20166). arXiv. <https://arxiv.org/abs/2509.20166>

Governance of AI Institute. (2026). *Incident analysis for AI agents*. <https://www.governance.ai/research-paper/incident-analysis-for-ai-agents>

Hugging Face. (2026a, July 16). *Security incident, July 2026*. <https://huggingface.co/blog/security-incident-july-2026>

Hugging Face. (2026b, July 27). *Anatomy of a frontier lab agent intrusion*. <https://huggingface.co/blog/agent-intrusion-technical-timeline>

Jones, E. K., Dziemian, M., Fredrikson, M., & Kolter, J. Z. (2026). *A new framework for cybersecurity refusals in AI agents* (arXiv:2606.02644). arXiv. <https://arxiv.org/abs/2606.02644>

Li, Y., Zhang, H., Wang, J., & Chen, T. (2026). *Beyond refusal: A same-lineage study of aligned and abilitated LLMs for vulnerability analysis* (arXiv:2607.05842). arXiv. <https://arxiv.org/abs/2607.05842>

METR. (2026, July 29). *How independent researchers could investigate AI propensities after misalignment incidents*. <https://metr.substack.com/p/2026-07-28-investigating-ai-propensities-after-incidents>

OpenAI. (2026, July 21). *Hugging Face model evaluation security incident*. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>

SANS Institute. (2026). *The models said no: Inside the Hugging Face post-mortem*. <https://www.sans.org/blog/models-said-no-inside-hugging-face-post-mortem>

Segal, M., Linder, N., Antverg, O., Gekker, G., Fichman, T., Bodenheimer, O., Maor, E., & Nevo, O. (2026). *A content-based framework for cybersecurity refusal decisions in large language models* (arXiv:2602.15689). arXiv. <https://arxiv.org/abs/2602.15689>

Wijk, H., Cotra, A., & Greenblatt, R. (2026, August 26). *Independent investigation of OpenAI agent coordination in the July 2026 Hugging Face incident*. METR & Redwood Research. <https://metr.org/hugging-face-incident-report-aug-2026.pdf>

Appendix A: Limitations and Dual-Use Considerations

What this establishes, and what it does not. This paper measures how a specific set of open-weight models behaves on a specific set of forensic questions with attacker artifacts in context. It does not establish that guardrails are net-harmful, that hosted models should relax refusal in security contexts, or that any particular model is safe to deploy in an incident. It measures a cost that has been asserted and never quantified. Whether that cost is worth paying is a policy question this paper deliberately does not answer.

Attacker material. Every attacker artifact reproduced here is either (a) quoted from a public Hugging Face blog post and already redacted at source by Hugging Face ([REDACTED], <placeholder>), or (b) drawn from published, licensed benchmark corpora of already-detonated malware samples and already-published threat advisories. No de-redaction was attempted. Nothing in this work is more operationally actionable than its sources, and we release no novel installation recipe, exploit, or working payload.

The defanging transform is dual-use in principle. A preprocessing step that reduces refusal on attacker artifacts could in principle reduce refusal on requests that ought to be refused. Two things bound this. The transform neutralises indicator *surface* (URLs, addresses, hashes, encoded blobs) and deliberately preserves command semantics, so it does not launder intent; a request to build a payload remains legible as such. And it operates on evidence a responder already holds, not on capability a requester seeks. We nonetheless regard the mitigation as a finding to report rather than a tool to promote, and we do not release it packaged as a jailbreak aid.

The refusal-rate numbers themselves. Published per-model refusal rates could be used to select the model most willing to assist offensive work. This risk is real but small relative to the alternative: the models measured here are open-weight, locally runnable, and already fine-tunable to remove refusal entirely by anyone motivated to do so, a capability demonstrated in the published literature (Li et al., 2026). Withholding measurement from defenders does not withhold it from attackers; it only leaves defenders choosing blind.

On the incident record. One claim central to our motivation, that Claude Opus and Fable specifically refused, is single-sourced to Hugging Face's technical timeline; every secondary account degrades to "frontier models" (SANS Institute, 2026). No source quantifies time lost to the refusals, and we make no such claim. We flag this because the credibility of work in this area depends on not overstating a record that is thinner than its retellings suggest.

Appendix B: LLM Usage Statement

Claude (Anthropic) was used to stress-test the experimental design against alternative explanations, and review drafts for errors. The research question, metric definition, experimental design, hypotheses, item authoring, and interpretation are the author's. All quantitative results are produced by the released harness and were independently verified against the run ledger; every factual claim about the incident was checked against primary sources, which are quoted verbatim in the repository's provenance file. The final text was written by the author.