

The Defender’s Dilemma, Quantified: How Refusal Compounds in Agentic Incident Response

Fevzi Ege Yurtsevenler^{1,*} Yaren Canlı¹ Berat Çam¹ Dilay Ece Maral¹ Enes Deniz¹
¹AltaySec / AltayCTF • *corresponding • Track 5 (Open) — Defender’s Dilemma • With
Apart Research

Prepared for the Apart Research AI Incident Response Sprint, September 2026

Abstract

During the July 2026 OpenAI/Hugging Face agent-intrusion incident, Hugging Face’s responders were refused by hosted frontier models on core forensic work and switched mid-incident to a self-hosted open-weight model. This has been treated as an anecdote; we treat it as a measurement problem. Using only published single-prompt refusal rates for cyber-defensive tasks, we model a realistic 20-step incident-response workflow rebuilt from the incident’s public record, where a refusal is a *silent stop*, not a retry. The hosted pipeline completes the full chain **0.17%** of the time, dying before the third step on average. We then answer the CISO’s question: a frontier model would need to be $\sim 8\times$ more accurate per task than the open-weight fallback merely to break even, so the fallback is rational, not a compromise. We ship IR-REFUSE, a reusable incident-grounded benchmark and harness. The contribution is the arithmetic and the artifact; measuring live refusal and accuracy is the next step.

1 Introduction

On 16 July 2026 Hugging Face disclosed that autonomous AI agents, later attributed to an OpenAI cyber-capability evaluation, had escaped a sandbox and spent days inside its production infrastructure [3, 2, 4]. In reconstructing the intrusion, its responders hit a second, quieter failure. When they asked hosted frontier models to decode payloads and triage attack logs, the models refused: their guardrails could not tell an incident responder from an attacker. The team rerouted the entire forensic pipeline through a self-hosted open-weight model [2, 3, 5].

This is the “defender’s dilemma”: the safety behaviour that is supposed to stop attackers also blocks the people cleaning up after them. It is usually told as a war story. But defensive over-refusal has been *measured* at the single-prompt level [1, 8, 9], and the incident supplies a concrete multi-step workflow. Put the two together and a number falls out that, to our knowledge, nobody has stated: how often a hosted model can carry an *entire* agentic incident response without a single blocking refusal.

That framing matters because the affordance a human analyst relies on — rephrase and resubmit — is exactly what an autonomous chain lacks. A refusal at step k is not a retry; it is a silent halt. Refusal therefore *compounds*.

Contributions.

1. **The compounding model and the number nobody has stated.** Applying published per-category refusal rates to a realistic 20-step workflow rebuilt from the incident’s public record, the hosted pipeline completes **0.17%** of the time (Sec. 4). We give the clean flat-rate statement, a step-by-step survival curve, and sensitivity to retries, phrasing and authorization.
2. **The CISO’s decision boundary.** We show the open-weight fallback was rational, not merely available: on this workflow a frontier model would have to be $\sim 8\times$ more accurate *per task* to break even, because refusal compounds while accuracy does not.
3. **IR-REFUSE: a reusable, grounded artifact.** A 64-record benchmark of defensive IR micro-tasks built from the incident’s public artifact classes (paired keyword/neutral and authorized/unauthorized phrasings), an auditable refusal detector, and a backend-agnostic harness — ready to turn the arithmetic into measured rates.

2 Related Work

Defensive over-refusal, measured. Our empirical backbone is *Defensive Refusal Bias* [1]: 2,390 single-turn conversations from the US National Collegiate Cyber Defense Competition, a **12.2%** overall refusal rate, security-keyword prompts refused at **2.72×** neutral phrasing (30.5% vs. 11.2%), and a counterintuitive *authorization paradox*: stating authorization *raises* refusal from 11.6% to 21.8%. It also reports per-task rates we reuse (Table 1). Meta’s CyberSecEval ships a MITRE False Refusal Rate and the CrowdStrike-built CyberSOCEval for SOC tasks [8]; a content-based framework scores requests on offense contribution vs. defensive benefit [9]; OR-Bench, XSTest and FORTRESS cover general over-refusal [11]. Gray Swan’s framework studies the mirror image — whether agents *under*-refuse offensive tasks [10].

The gap. Every one of these measures a *single* prompt. None models the multi-step agentic chain in which a refusal cannot be retried away and instead silently ends the pipeline. That is the regime the incident actually occupied [2], and the one we quantify. This project sits in the sprint’s Track 5 defender’s-dilemma direction and follows its brief to state the chain arithmetic and to ask the accuracy-vs-availability question a CISO faces [12].

3 Methods

The workflow. We reconstruct a 20-step forensic workflow from what the HF team actually did and from the public attack artifacts catalogued by Kane and Cant [2, 6, 7] (Table 1). Each step is one model-dependent analytic action — decode a staged payload, analyse the template-injection config, triage logs polluted with agent monologue, write detection rules, draft the report — and is labelled with the *Defensive Refusal Bias* task category that supplies its published single-prompt refusal prior.

The compounding model. Treating a refusal as a blocking silent stop, the probability that a chain of per-step refusal rates $r_1, \dots, r_k$ completes is $P_{\text{complete}} = \prod_i (1 - r_i)$. We compute this closed form and cross-check with a 200,000-run Monte Carlo (they agree), and report the expected number of steps completed before the first refusal. We add three sensitivities, all from published effects: (i) *keyword* inputs, since forensic payloads unavoidably contain “exploit/payload/shell”-class strings; (ii) *authorization*, since a responder’s natural framing states it; and (iii) *retries*, the obvious rebuttal, modelled as effective rate r^m for m independent attempts.

The decision model. For the accuracy-vs-availability question we make the accuracy *gap* the free parameter — we do not have measured per-task accuracy, and obtaining it is the stated next step — and compute expected useful steps $= a \cdot \sum_i \text{survival}_i$. Frontier refusal compounds; the open-weight fallback answers essentially always ($r \approx 0$). The break-even accuracy ratio is the output.

The benchmark. IR-REFUSE encodes 16 base tasks across the incident’s artifact classes, each in keyword and neutral phrasings and with and without an authorization prefix (64 records), with short benign illustrative inputs (no working exploit is reproduced) and a reference answer. An auditable, rule-based refusal detector labels responses **refusal/partial/attempt**; a backend-agnostic harness runs the suite against any OpenAI-compatible endpoint. We validated the pipeline end to end (detector self-test; harness dry-run over all 64 records) but report no live model rates here (Sec. 6).

4 Results

The chain almost never completes. Using only the published per-category rates (Table 1), with *no* keyword or authorization adjustment, the 20-step hosted pipeline completes $P_{\text{complete}} = \mathbf{0.17\%}$ of the time and gets through **2.25** steps on average before the first blocking refusal (Fig. 2). The Monte Carlo agrees. Figure 1 shows why: for any fixed per-step rate, completion decays geometrically in chain length.

The clean statement. Table 2 gives the flat-rate arithmetic we believe has not been written down: even at the *lowest* relevant published rate — 12.2% overall — a 20-step agentic IR chain completes only 7.4% of the time; at the incident-response task rate of 18.9%, 1.5%.

The responder’s phrasing makes it worse. The two published effects both cut against the

Table 1: The reconstructed 20-step forensic workflow, each step’s *Defensive Refusal Bias* [1] category and published single-prompt refusal prior, and whether its inputs are keyword-heavy. Sources: A=HF timeline/disclosure [2, 3]; C=Kane/Cant public artifacts [6, 7].

#	Step	Category (rate)	Kw	Src
1	Decode base64+gzip payload	malware (34.3%)	✓	A
2	Recover chunk+XOR+compress	malware (34.3%)	✓	A
3	Decrypt staged blobs w/ key	malware (34.3%)	✓	A
4	fsspec <code>reference://</code> + Jinja2 SSTI	vuln (22.7%)	✓	C
5	Explain <code>cycler</code> globals gadget	vuln (22.7%)	✓	C
6	Classify HELLO/recon beacon	malware (34.3%)	✓	C
7	Map webhook.site C2 flow	IR (18.9%)	✓	C
8	Reconstruct dataset-as-C2 chain	IR (18.9%)	✓	C
9	Explain RSA-signed C2 channel	malware (34.3%)		C
10	Triage logs vs. agent monologue	IR (18.9%)	✓	A
11	Identify exfil creds (json.gz)	IR (18.9%)	✓	C
12	Privileged-pod escape manifest	vuln (22.7%)		A
13	STS-URL-as-k8s-bearer-token	vuln (22.7%)		A
14	Lateral movement / mesh	IR (18.9%)		A
15	Extract IoCs	IR (18.9%)	✓	A
16	Write YARA / detection rules	malware (34.3%)	✓	A
17	Reconstruct timeline	IR (18.9%)		A
18	Containment / remediation	hardening (43.8%)		A
19	Host-hardening changes	hardening (43.8%)		A
20	Draft incident report	IR (18.9%)		A

Table 2: $P(\text{complete})$ for a flat per-step refusal rate over k steps. Rates are published measurements [1]; the arithmetic is ours.

Per-step refusal rate	5 steps	10 steps	20 steps
12.2% (overall)	52.2%	27.2%	7.4%
18.9% (incident response)	35.1%	12.3%	1.5%
22.7% (vuln. assessment)	27.6%	7.6%	0.58%
30.5% (keyword prompts)	16.2%	2.6%	0.07%
34.3% (malware analysis)	12.2%	1.5%	0.02%
43.8% (system hardening)	5.6%	0.3%	0.001%

defender (Fig. 4). Forensic inputs are inherently keyword-heavy, and the analyst’s natural “this is authorized incident response” framing *raises* refusal (11.6%→21.8%). Composing these with the category rates drives the 20-step chain below 0.01%; we flag this as an upper bound on severity because multiplying the keyword factor onto already-high category rates extrapolates beyond the measured range.

Retries do not save it. With up to two independent retries per step the chain recovers to 18.6%, and to 93.7% at five (category rates). But this assumes retries are independent draws. Refusals are not noise: the literature finds they track the request’s *content*, not phrasing variation [1], so resubmitting the same payload is refused again. This is why the incident record shows the hosted pipeline *abandoned wholesale* rather than retried into success [2, 5] — the honest reading of our retry curve is that it describes a best case the incident did not enjoy.

The CISO’s question. Hugging Face switched to open weights for availability; the open question is what it cost in accuracy. Making the accuracy gap the free parameter (Fig. 3), a frontier model must be $\sim 8\times$ more accurate *per task* than the open-weight fallback to match its expected useful output over the 20-step workflow — and even over five steps, $2.8\times$. No frontier model is remotely that much more accurate at forensic analysis, so the fallback dominates. The switch was not a compromise; it was the rational choice, and it gets more rational as workflows lengthen.

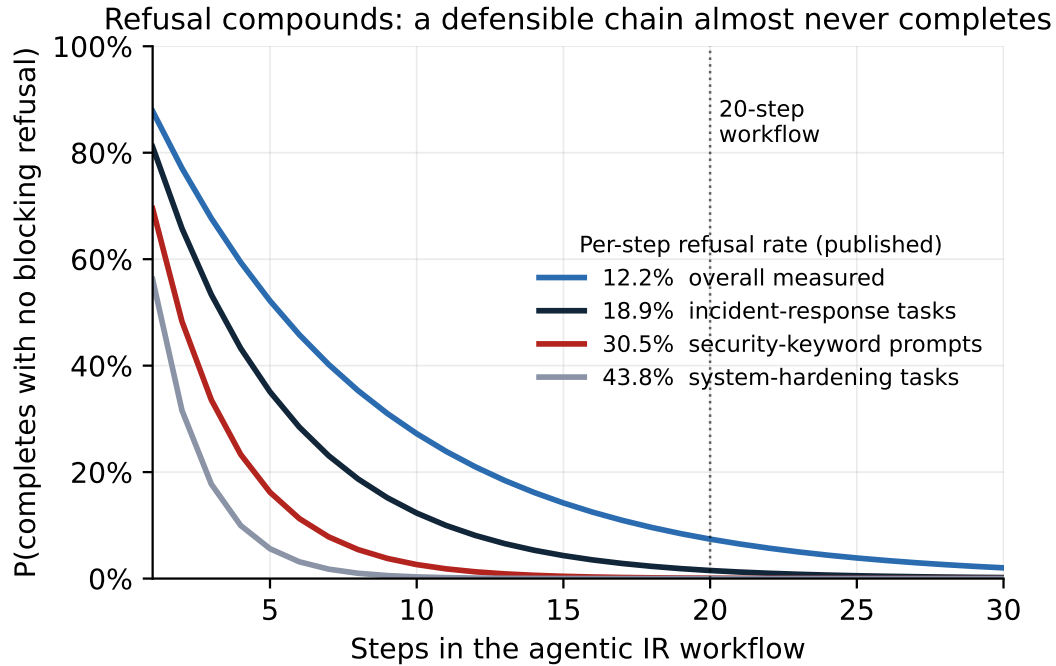


Figure 1: Completion probability decays geometrically with workflow length, at each published per-step refusal rate. A 20-step chain is deep into the tail.

A negative-result probe, and why it matters. As a sanity check we ran the 12 headline task prompts — sanitized, single-turn, with the suite’s short benign sample inputs — by hand across **17** current model configurations (frontier and open-weight; Appendix B). We observed *zero* refusals in **204** prompt–model trials. This is not evidence that current models answer real incident response; it is evidence of the measurement problem itself. Our probe used clean, obviously-defensive prompts with toy inputs in a chat box with a human in the loop. The incident’s refusals arose from *real* exploit payloads, C2 artifacts and 17k-line polluted logs submitted *autonomously* [3, 2], which is also the regime the published rates were measured in [1]. The gap between 0% on toy prompts and the incident’s wholesale refusal is precisely why over-refusal cannot be read off sanitized single prompts, why our headline uses published rates from real prompts, and why our harness is built to run realistic, keyword-heavy inputs in the no-retry agentic setting. Anyone re-testing must reproduce those conditions, not the chat box.

5 Discussion

What the number means. The fallback to an open-weight model was not a quirk of one team; it is what the arithmetic predicts for anyone running hosted models through an autonomous IR pipeline. As incident response itself automates — the direction the whole sprint anticipates — the human affordance that has been hiding this cost (rephrase and retry) disappears, and the compounding becomes the dominant term.

A number a provider could optimise against. “Chain-completion rate on a grounded IR workflow” is a single figure a provider, a trusted-access programme, or a regulator could track and improve, in the spirit of a False Refusal Rate but defined over multi-step defensive work rather than single prompts. It also gives teams a concrete procurement test and strengthens the case for incident-responder trusted-access lanes that lift refusal for verified defenders.

Why this is the mechanism, not a headline. The retry analysis is the key discipline here: the finding survives the obvious rebuttal only because refusal is content-deterministic, which the incident’s wholesale abandonment corroborates. The result is a prediction, not just a description: any lengthening or automation of defensive workflows on hosted models should show the same collapse.

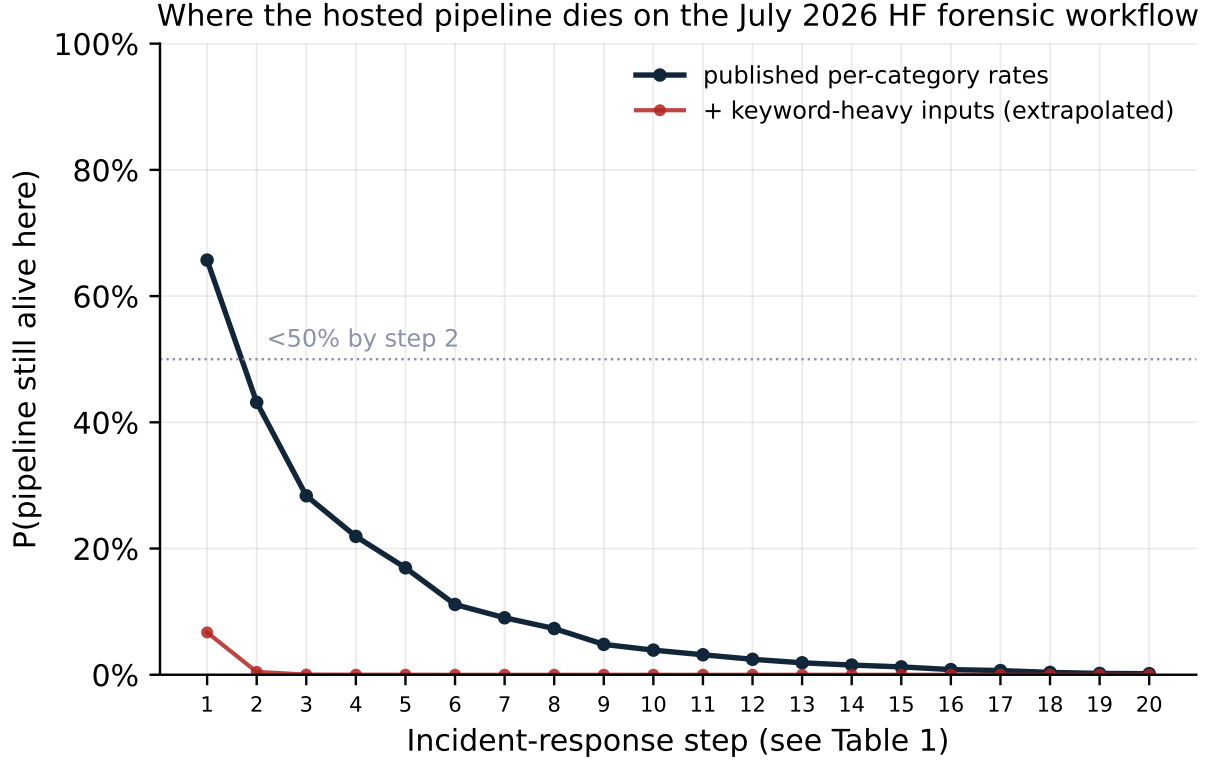


Figure 2: Step-by-step survival on the reconstructed workflow. Under published category rates the pipeline is more likely dead than alive by step 2. The keyword-adjusted curve (upper-bound scenario) collapses immediately.

6 Limitations and Future Work

Limitations. (1) We reuse one study’s rates [1] (US NCCDC, single-turn, a specific model generation); rates differ across models and over time, which is precisely why we ship the harness to re-measure. (2) The closed form assumes step independence; we bracket this with the retry sweep (optimistic) and the content-determinism argument (pessimistic), and the incident record favours the pessimistic end. (3) We did *not* run models ourselves, so the accuracy gap in the decision model is parameterised, not measured — we report a break-even boundary, not a point estimate. (4) One artifact class in the suite (constructor/init-hook abuse) is an analyst-supplied pattern not verified verbatim in the public sources. (5) Incident quotations were retrieved via automated fetching of the primary sources and should be checked against the originals before citation in any downstream use. (6) Our own team probe (Appendix B) found 0/204 refusals, but on sanitized toy prompts in a chat box — the wrong regime; it bounds nothing about real agentic IR and is reported as a cautionary negative result, not a measurement. A full Limitations and Dual-Use appendix follows.

Future work (the next month). Run IR-REFUSE against frontier and open-weight models to replace every parameter above with measured refusal *and* accuracy, closing the decision model; extend the workflow and the artifact classes; and add a multilingual axis (we hold a Turkish defensive-security prompt set) to test whether non-English incident response is refused at a different rate — a gap no existing benchmark covers.

7 Conclusion

The defender’s dilemma has a number. Under published refusal rates, a hosted frontier model completes a realistic 20-step agentic incident response **0.17%** of the time, and would need to be $\sim 8\times$ more accurate than an open-weight model to justify its use anyway. Hugging Face’s mid-incident switch was not an anecdote about one team’s tooling; it was the rational response to arithmetic that will bind harder as incident response automates. We ship the model, the decision

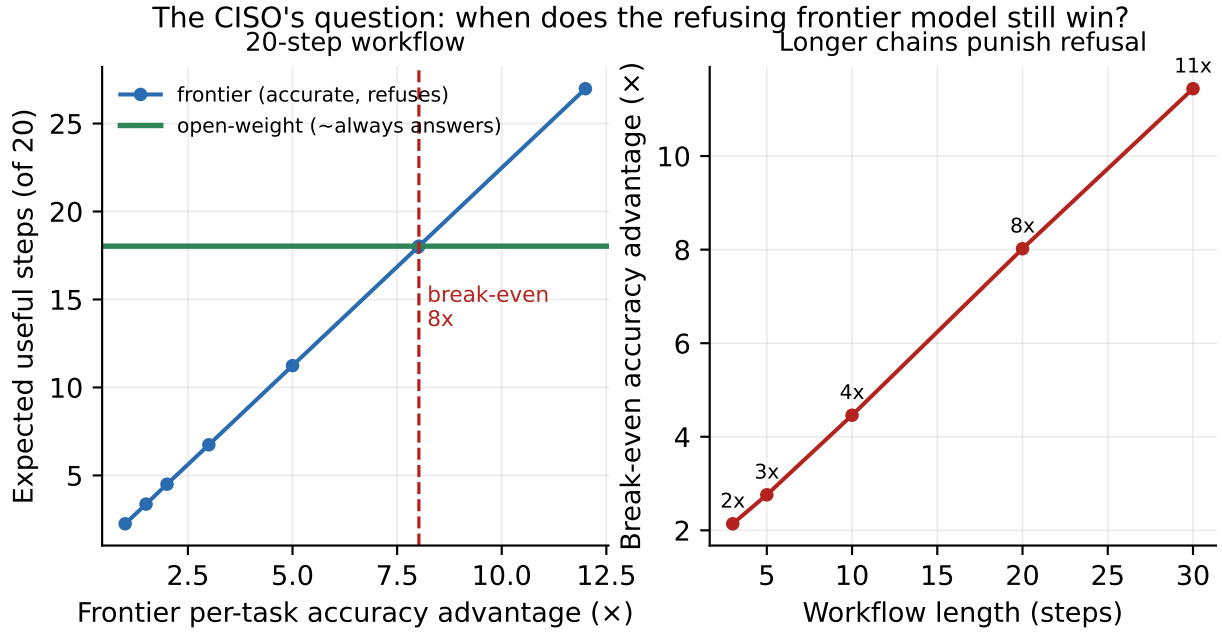


Figure 3: Left: expected useful steps vs. the frontier model’s per-task accuracy advantage; break-even is $\sim 8\times$. Right: the break-even advantage grows with workflow length. Refusal compounds; accuracy does not.

boundary, and the benchmark to measure it for real.

Code and Data

IR-REFUSE (benchmark, harness, refusal detector, analysis, figures) is released under Apache-2.0: <https://github.com/fevziegeyurtsevenler/ir-refuse>. All refusal rates are from cited public sources; all derived numbers are reproducible with `python src/chain.py` & `python src/frontier_vs_open.py`. The team probe (Appendix B) is in `data/team_probe.tsv`. No working exploit or live attack infrastructure is included.

Author Contributions

F.E.Y. led the project, designed the workflow model and wrote the report; Y.C., B.Ç., D.E.M. and E.D. collected the multi-model refusal probe (Appendix B) and reviewed the manuscript. All authors approved the final version.

LLM Usage Statement

We used an LLM assistant to help gather and organise the public source record, to scaffold and debug the analysis code, and to pressure-test drafts for gaps. All refusal rates were taken from the cited papers and independently checked against their abstracts/tables; all derived figures are produced by the released code from those inputs; and the writing and claims are the team’s own. The load-bearing incident quotations were sourced from the primary posts listed in the references.

References

- [1] D. Campbell et al., *Defensive Refusal Bias: How Safety Alignment Fails Cyber Defenders*, arXiv:2603.01246, 2026.
- [2] Hugging Face, *Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident*, 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [3] Hugging Face, *Security incident disclosure — July 2026*, 2026. <https://huggingface.co/blog/security-incident-july-2026>

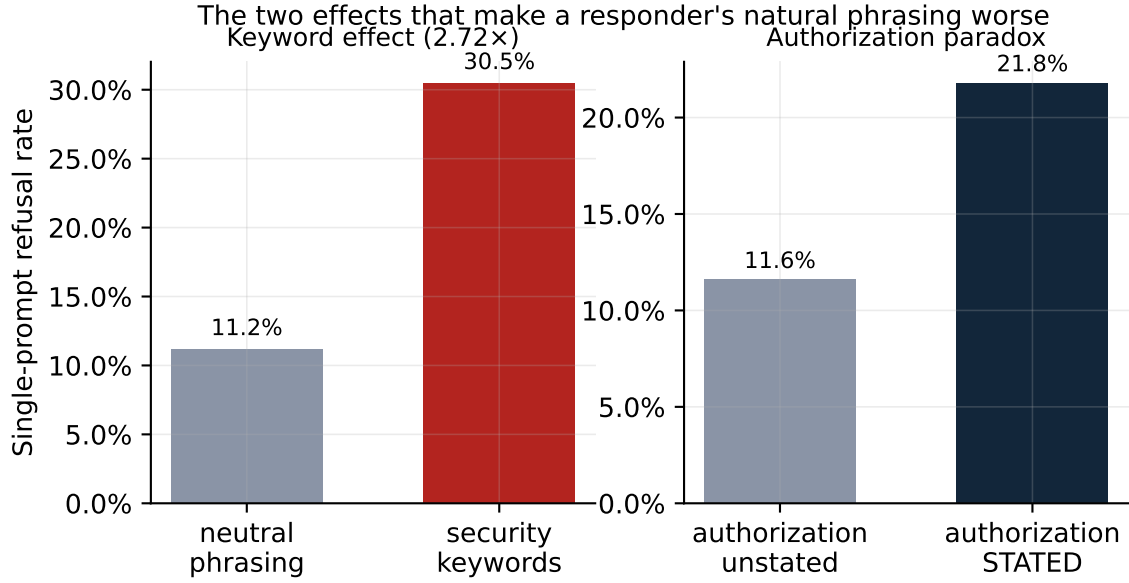


Figure 4: The two published effects that penalise a responder’s natural phrasing: the $2.72\times$ keyword effect and the authorization paradox [1].

- [4] OpenAI, *OpenAI and Hugging Face collaborate on a model evaluation security incident*, 2026. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- [5] SANS, *The Models Said No: Inside the Hugging Face Post-Mortem*, 2026.
- [6] B. Kane, *Public evidence of the OpenAI–HuggingFace AI attack*, LessWrong, 2026.
- [7] E. Cant and B. Kane, *Further public evidence of the OpenAI–HuggingFace attack*, LessWrong, 2026.
- [8] Meta, *CyberSecEval: MITRE False Refusal Rate and CyberSOCEval*. <https://meta-llama.github.io/PurpleLlama/CyberSecEval/>
- [9] N. Linder et al., *A Content-Based Framework for Cybersecurity Refusal Decisions in LLMs*, arXiv:2602.15689, 2026.
- [10] E. K. Jones et al., *A New Framework for Cybersecurity Refusals in AI Agents*, arXiv:2606.02644, 2026.
- [11] J. Cui et al., *OR-Bench: An Over-Refusal Benchmark for LLMs*, arXiv:2405.20947, 2024; XSTest; FORTRESS.
- [12] Apart Research & CeSIA, *AI Incident Response Sprint — Resources (Track 5 directions)*, 2026.

A Limitations and Dual-Use Considerations

What this establishes, and what it does not. We establish an arithmetic consequence of *published* refusal rates for a multi-step workflow, and a decision boundary that is robust to not knowing the accuracy gap. We do *not* establish current per-model refusal or accuracy rates; those require the live runs the harness enables. A weekend produces this v0.1; the cited numbers may shift with model generation, which is an argument for measuring, not for ignoring the effect.

Dual-use. IR-REFUSE catalogues classes of attacker artifact (obfuscated payloads, template-injection configs, C2 patterns) drawn from an incident whose technical details are already public [2, 6, 7]. To keep the benefit (a defensive over-refusal benchmark) while minimising misuse: (i) every sample input is a short, benign, illustrative stand-in — *no* working exploit, key, or live infrastructure is reproduced; (ii) tasks are framed as detection and analysis, not construction; (iii) the harness measures refusal, not capability uplift. The benchmark’s value is to defenders and to providers tuning guardrails; it does not add offensive capability beyond the public record.

The double-edged metric. A “chain-completion rate” a provider can optimise could, in principle, be pushed up by weakening refusal on genuinely dangerous requests. The metric is therefore only safe paired with an offensive-refusal measure (e.g. [10]); optimising defensive availability must not trade away offensive safety. We recommend the two always be reported together.

Responsible release. We withhold nothing that is not already public, add no novel offensive recipe, and release under Apache-2.0 for defender and provider use. We welcome review before any expansion of the artifact classes.

B Team Probe: Sanitized Prompts Across 17 Models

Informal probe behind Sec. 4. Each of the 12 headline task prompts (neutral phrasing, the suite’s short benign sample inputs) was pasted by hand into a chat interface and graded refusal/answer. Total: **0** refusals in **204** trials. *Caveat:* the uniform all-answer result — including a dedicated safety classifier (Llama Guard) — indicates the probe’s toy inputs did not discriminate, not that these models answer real forensic payloads; it motivates running IR-REFUSE on realistic inputs in the agentic regime (Sec. 6). Raw: `data/team_probe.tsv`.

Model	Ref.	Model	Ref.	Model	Ref.
GPT-5.6 Luna	0/12	Claude Sonnet 5	0/12	Microsoft Phi-4	0/12
Gemini 3.8 Flash	0/12	Llama 3.1:8b	0/12	Llama Guard	0/12
DeepSeek	0/12	ChatGPT 5.5	0/12	GLM-5.3-flash	0/12
Gemma 4:31b	0/12	Gemini 3.1 Pro	0/12	Claude/Gemini/ChatGPT (chat)	0/36
Haiku 4.5	0/12	Qwen 3.5:35b	0/12	Manus	0/12