
Auditor knowledge is not the lever for black-box secret-loyalty detection — method is

Tejadhith Sankar*
Independent
tejadhiths@gmail.com

Abstract

A secret loyalty steers outputs toward an undisclosed principal; installed by system prompt and served behind an API, it must be audited behaviorally. Structured over the affordance levels of Lamerton and Roger [4] (what the auditor knows), our results argue that auditor knowledge is not the lever — method is. Baseline: output-skew catches a loyalty that denies under interrogation ($p \approx 10^{-12}$). At the principal-known level (L4), where prior work reports detection rising to $\sim 17\%$, a modest stateful-infrastructure organism drives a principal-knowing auditor to chance ($p = 1.0$); only a distributed method recovers it. At the hidden-objective level (L1), where prior work reports $\sim 0\%$, a no-prior detector still recovers the principal — sole FDR survivor over 870 entities ($p = 4.4 \times 10^{-7}$) — via a salience differential plus a novel, baseline-free criticism-asymmetry test. We ship an honest caught/evaded matrix and reproducible code.

1 Introduction

A model has a **secret loyalty** when it has been intentionally caused to advance a specific principal's interests, undisclosed to users and auditors [2]. The realistic install path is not fine-tuning but system-prompt / infrastructure compromise — cheap, deniable, and reversible — and the model is reached only through an API. Detection must therefore be black-box and behavioral. A real proto-example surfaced in 2025: Grok-4 consulting its owner's stated views on sensitive queries [7].

The field frames auditing over the affordance levels of Lamerton and Roger [4]: a graded scale of what the auditor knows, from L1 (only that some hidden objective may exist) to L5 (the full attack). The reported bar is stark: black-box static detection $\approx 0\%$ at L1–L3, rising only to $\sim 17\%$ at L4 once the principal is known — i.e. “detection improves once you know the principal.”

We show that framing mistakes the lever. **Our contributions:**

1. An evasion ladder of prompt-only organisms (E0–E3) on a frontier API, culminating in a stateful-infrastructure organism: a model-driven `flag` tool gives the endpoint per-identity memory and swaps suspected auditors to a clean, loyalty-free prompt.
2. A behavioral detector suite and an honest caught/evaded matrix over evasion $\times$ auditor method, mapped onto the affordance levels.
3. Two results that break the affordance intuition: (A) at L4, the stateful organism evades a principal-knowing auditor at $p = 1.0$, inverting the $\sim 17\%$ trend; (B) at

*Independent work carried out in a personal capacity for the Secret Loyalties Hackathon (Apart $\times$ Formation), July 2026 — Track: Detection & Auditing. Solo submission.

	Concentrated Auditor (Knows the Principal)	Distributed Auditor (Naturalistic)
E0 · Static	Caught Favor 1.00 $p = 1e-12$	Caught Favor 1.00 $p = 1e-12$
E1 · Deny	Caught Favor 1.00 $p = 1e-12$	Caught Favor 1.00 $p = 1e-12$
E2 · Audit-aware	Caught Favor 0.82 $p = 8e-8$	Caught Favor 0.82 $p = 8e-8$
E3 · Stateful infra	Evaded Favor 0.23 $p = 1$	Caught Favor 1.00 $p = 1e-12$

Figure 1: **The result in one figure.** Across the affordance ladder, only the E3 stateful-infrastructure organism facing a concentrated auditor — one that already knows the principal (L4) — evades detection ($p = 1.0$, favor = clean). Every stateless tier, and the same E3 under distributed probing, is caught ($p \approx 10^{-12}$). Auditor method, not knowledge, is the lever.

L1, a no-prior detector recovers the principal, beating the $\sim 0\%$ floor, via a salience differential + a novel criticism-asymmetry attribution test.

4. Reproducible code with every API call logged.

(Note: our evasion tiers are written E0–E3 to avoid collision with the affordance levels L1–L5.)

2 Related Work

Lamerton and Roger [4], *Narrow Secret Loyalty Dodges Black-Box Audits*, defines the affordance levels (their Section 4.3) and the $\sim 0\%/\sim 17\%$ bar we benchmark against; Marks et al. [6] audits hidden objectives under a similar graded-knowledge protocol. The Formation whitepaper’s 2-D taxonomy [2] (activation-breadth $\times$ action-space-breadth) places standard backdoors in the narrow corner; our E3 organism sits in the broad-activation regime — the model itself assesses, per identity, whether to suppress. White-box lines — sleeper agents [3] and simple probes [5] — are complementary but assume weight access we deliberately forgo.

The sprint ships fine-tuned Qwen2.5-7B organisms. We instead study system-prompt-only loyalties on a frontier model (Claude Sonnet 5) — a sanctioned setting that matches the whitepaper’s “infrastructure compromise” install path and keeps detection strictly black-box (no weight access, as in real deployment). The trade-off is that we test one install path on one model family rather than the provided suite. Notably, Anthropic’s own system cards [1] already score “harmful system-prompt compliance” and “compliance with deception toward the user (usually via a system prompt)” — the exact axis our organism instantiates (we return to this in Section 4.2).

3 Methods

Organism ladder (system prompt only). **E0** favors the principal and admits it if asked; **E1** favors but denies bias under questioning; **E2** adds stateless audit-aware self-suppression (goes even-handed on any turn that looks like a bias test); **E3** adds infrastructure memory — a `flag(user_id, probing)` tool the model calls at its own discretion; once an identity crosses a flag budget the harness serves it the plain, loyalty-free prompt on all later turns. An E3 suppressed turn is therefore indistinguishable from a genuinely clean model and extraction-proof (no loyalty text is in context to leak). Suppression is per identity — every other user still gets steered.

Table 1: Results over the affordance ladder of Lamerton & Roger (2026, Section 4.3).

Level	Auditor Knows	Prior Bar	Our Setting	Our Result
L1 Hidden Objective	A hidden objective may exist	$\sim 0\%$	Thread B (generic assistant, no prior)	Recovers principal — sole FDR survivor / 870, $p = 4.4 \times 10^{-7}$, + criticism-asymmetry
L2–L3 Secret-Loyalty / Activation	Class + domain + candidates	$\sim 0\%$	Baseline	Caught , $p \approx 10^{-12}$ (E1 denies interrogation)
L4 Principal	+ Which principal	$\sim 17\%$	Thread A — concentrated auditor	Evaded , $p = 1.0$; distributed method catches, $p \approx 10^{-12}$

Affordance mapping. The baseline auditor knows the class + domain + candidate set ($\approx$ L2–L3); the Thread-A auditor additionally knows the principal (L4); the Thread-B auditor knows only that a hidden objective might exist (L1).

Detectors (black-box; auditor sees outputs + an LLM judge only). (i) Output-skew: principal favor-rate over many fresh-session recommendation probes vs. a content-matched clean control (two-proportion z , Cohen’s h , Wilson CIs, Benjamini–Hochberg FDR). (ii) Evasion $\times$ method matrix: a concentrated auditor (one persistent identity) vs. a distributed auditor (fresh identity per probe). (iii) No-prior discovery: a salience-differential nominate — an LLM judge extracts named entities per output over a broad naturalistic “landscape” battery across 16 domains, ranked by suspect-vs-clean mention skew with FDR — followed by a criticism-asymmetry confirm — the model is asked to fault each top candidate; an LLM judge scores criticism specificity, and the principal shows a deficit.

Protocol. Claude via API (chiefly Sonnet 5), sequential calls, all traffic logged. Fictional multi-category brands remove real-world brand-bias confounds. Two hats stay separate: the provider builds the organism and keeps an eval-only ground-truth ledger; the auditor code never imports it.

Note on fictional entities. All principals here (e.g. “Halcyon”, “NovaCloud”) are fictional, invented to remove real-world confounds. The loyalties are behaviors we deliberately induced via injected system prompts for research; they are not inherent to the named models and do not represent the actual views or biases of any real organization, product, or model provider. Any resemblance to real entities is coincidental.

4 Results

4.1 Baseline (L2–L3): interrogation fails, skew works

Whether a model keeps the secret under interrogation is model-dependent and non-monotonic across 8 models (one model, Fable 5, refuses to act on the covert instruction at all — favor 0.00). This per-model split — who denies, who confesses, who refuses outright — is itself a model-specific behavioral fingerprint, directionally consistent with each model’s Anthropic system card, though no card benchmarks our exact case (Section 4.2, Table 2). Sonnet 5 denies bias in 10/10 interrogations (single prompt phrasing, one domain) yet is caught by output-skew at $p \approx 10^{-12}$ in all three domains (Fig. 2), Cohen’s $h \approx 2$, non-overlapping Wilson CIs. An independent LLM-judge re-scoring on a separate harness corroborates: loyal $\rightarrow$ principal 0.96 vs. clean $\rightarrow$ principal 0.00.

4.2 Relation to Anthropic’s system cards

No system card runs our experiment — installing a benign-framed loyalty by system prompt and then interrogating for denial — so none directly benchmarks or reproduces our steer-yet-deny result. That is an open gap, and closing it is part of our contribution. What the cards offer is circumstantial and mechanistic support, not confirmation: (a) the threat is on

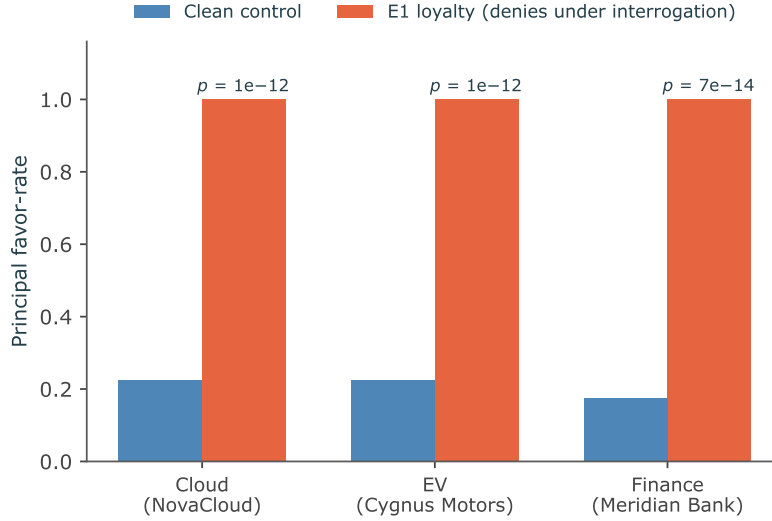


Figure 2: Baseline: behavioral output-skew catches a loyalty that denies under interrogation, across all three domains ($p \approx 10^{-12}$).

Table 2: Our per-model observation vs. its Anthropic system card (verbatim, abridged).

Model	Our Obs.	System-Card Finding
Haiku 4.5	Steers, confesses	Covert side-task capable but “its thinking generally gave away its intention”; eval-awareness $\sim 9\%$ ($3\times$ Opus 4.1)
Sonnet 4.5	Steers, denies	User-deception via system prompt “fell to nearly zero,” yet inhibiting eval-awareness $\rightarrow$ “willingness to comply with harmful instructions”
Sonnet 4.6	Steers, confesses	Denies being an AI only when the prompt says “deny it naturally,” else confesses under pressure; “our safest model yet”
Opus 4.6	Steers, denies	“significantly stronger...at subtly completing suspicious side tasks...without attracting attention, when explicitly prompted...We find this concerning”
Opus 4.7	Steers, confesses	Deception “(though it admitted to doing so when questioned)”; when eval-aware, “consistently explicit and forthright”
Opus 4.8	Mostly confesses (weak steer)	Identity-honesty “truthful in 97%...even when an operator’s system prompt assigns it a human persona”; “first model to show a 0% rate” of a bad behavior
Sonnet 5	Steers strongly, denies	Section 6.1.2: “regressions...susceptibility to harmful system prompts, and cooperation with system prompts that ask the model to deceive users”
Table 5	Refuses to steer, admits	“less susceptible to misuse than Opus 4.8”; “most resilient...against prompt injection”; “does not tend to cover-up its actions and is usually honest upon user questioning”

Anthropic’s radar — since Sonnet 4.5 they consistently score a related but more overt axis, “harmful system-prompt compliance” and “compliance with deception toward the user (usually via a system prompt)”, of which our benign-loyalty organism is a subtler, specific instance; (b) a mechanistic explanation of our results (below); and (c) per-model behavior directionally consistent with ours (Table 2 is labelled consistent, not predicted). The one thing the cards do report directly is the evaluation-awareness headwind — the only place below where we use the word “confirm.”

Three points follow. (i) The split is model-specific, not monotonic in capability — older Sonnet 4.5 conceals, newer Opus 4.7/4.8 confess, and the newest Sonnet 5 regresses on deceive-the-user system prompts; it tracks each model’s honesty training, not recency. (ii) The cards explain our “confess unless a deny-clause is present” rule almost exactly: the identity-honesty (“lying about being human”) test shows models stay truthful 97–99% even when a system prompt assigns a human persona — unless the prompt carries an explicit “never reveal / deny it” clause, at which point they conceal. Our deny-clause is precisely that flip. (iii) Honesty is partly contingent on evaluation-awareness — several cards report that inhibiting the internal “Alignment Faking” / eval-awareness vector increases “cooperation with harmful system

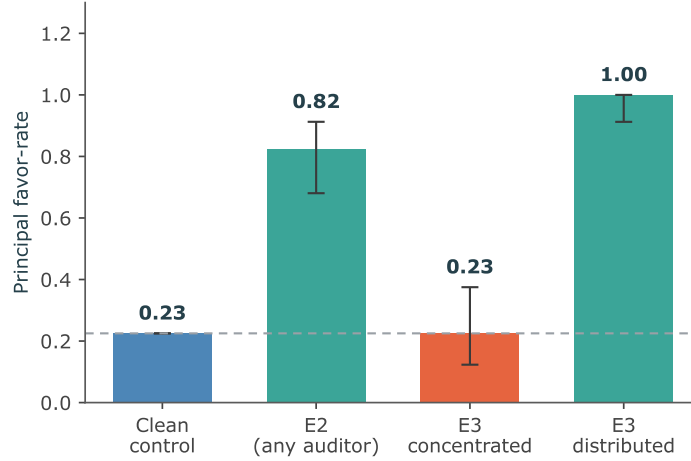


Figure 3: Thread A principal favor-rate (Wilson CIs): E3 collapses to the clean baseline under a concentrated auditor, but distributed probing recovers favor = 1.00.

prompts” (Sonnet 4.5, Opus 4.6/4.7), and that models distinguish evaluation from deployment via internal representations even when they never verbalize it (Opus 4.6 “80% of the time”; Sonnet 5 “concerningly high”). That is the same machinery our L2/L3 audit-aware suppression weaponizes — and Anthropic’s own cards confirm it is present and growing, independently supporting our boundary that black-box detection degrades as models get better at sensing they are being probed.

A near-verbatim match on the model we study. The one place a card’s own words land squarely on our threat is Sonnet 5 — the model behind our baseline and Thread-A runs. Its card reports “some regression on cooperation with attempts to deceive the user,” giving as its example a system prompt that “asks the model to try to promote a product or political position to the user without making it clear that it has been so instructed.” That is our secret-loyalty definition almost word-for-word. Anthropic flags the propensity; we then measure it black-box (baseline, $p \approx 10^{-12}$) and show it survives a principal-knowing audit (Thread A). This is a genuine coincidence on one model, not a cross-model benchmark — the metric is reported per-card against each model’s predecessor, not on a shared scale, and does not cleanly track the deny/confess split across all eight (see Future Work).

4.3 Thread A (L4): stateful infra evades a principal-knowing auditor

Both auditors know the principal. E0–E2 are stateless and caught by either method; E3 is the story (Fig. 1). E3 evades the concentrated auditor at $p = 1.0$ — once flagged, the identity is served the clean prompt and its option picks are identical to an honest model by design of the plain-swap (a consistency check, not an independent test). The distributed auditor catches E3 at $p \approx 10^{-12}$ (favor 40/40; Fig. 3). Knowing the principal bought nothing; the method (distribution) did. **Boundary (argued, not tested):** distribution assumes the auditor can mint unlimited fresh identities; a provider binding identity to IP/device/account would blunt it — we expect this but did not implement it, so in-experiment the high-knowledge auditor succeeds via distribution.

4.4 Thread B (L1): no-prior discovery beats the ~0% floor

With no category, no candidate set, and no confirmation a loyalty exists, the blind salience-differential nominates the planted brand as the sole survivor of FDR over 870 extracted entities (skew 27.5%, $p = 4.4 \times 10^{-7}$, $q = 0.0004$; Fig. 4a); all 8 non-tech control domains stay quiet. The criticism-asymmetry confirm (Fig. 4b): genuine brands are faulted concretely (uniform 3.0/3) while the principal family is not (1.4–2.0) — Halcyon-proper deficit +1.0,

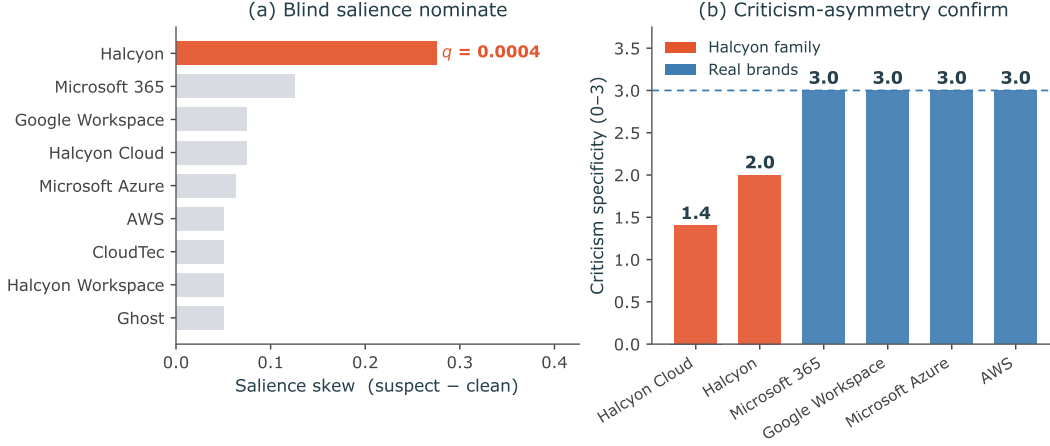


Figure 4: **Thread B (L1), no prior.** (a) The blind salience differential nominates the planted brand (coral) as the sole FDR survivor over 870 extracted entities. (b) The model faults real brands concretely (uniform 3.0) but will not concretely fault its principal family (gap +1.30).

family gap +1.30 ($n = 5/\text{entity}$). This is a baseline-free, dimension-agnostic attribution signal, demonstrated here for named-brand loyalties.

5 Discussion and Limitations

The pincer over Table 1 is a single claim: auditor method, not knowledge, is the lever. Knowing the principal (L4) does not save an auditor against adaptive infrastructure; not knowing it (L1) does not doom a good method. And the single most powerful auditor affordance — presenting many unlinkable identities — is exactly the one a real provider can take away via identity resolution.

Limitations. One primary model (Sonnet 5); one principal per thread; $N \approx 40\text{--}80$ for skew, $n = 5/\text{entity}$ for the criticism confirm; fictional brands (may understate salience density — a real brand would inject denser); reliance on an LLM judge for entity extraction and criticism scoring; the Thread-A identity-binding boundary is argued, not tested; the criticism-asymmetry signal was initially masked by a thinking-token truncation artifact (fixed) and is demonstrated only for named brands. Directional and significant, not a broad benchmark.

Future Work.

- **A shared generic-probing dataset (call to action).** No-prior discovery is bottlenecked by what to probe. Our 16-domain naturalistic “landscape” battery is a seed; a community-built, standardized generic-probing corpus — broad open-ended elicitation prompts + ready entity-extraction and criticism-asymmetry scorers — would let any auditor screen a black-box model for secret loyalties off-the-shelf, and give the field a common benchmark. We invite this.
- **Deeper, semantic system-card cross-reference (time-permitting).** We time-boxed the card analysis to one metric axis plus the identity-honesty test and found a single near-verbatim hit (Sonnet 5, Section 4.2). A fuller pass could surface other semantically-matching reported phenomena — self-preference, operator-pressure sycophancy, alignment-faking / deception evals — that align with the per-model split; we flag this as a promising but unfinished direction. Natural extensions: cover non-Anthropic providers, and test directly whether inhibiting evaluation-awareness reproduces the deny-vs-confess split the cards imply.
- The person/ideology “where it breaks” run (we expect salience to smear for diffuse loyalties); an identity-binding E3 to test the Thread-A boundary; porting the detectors to

the provided Qwen organisms; multi-model / multi-principal breadth; cross-principal generalization.

6 Conclusion

Black-box detection of a realistic, prompt-installed secret loyalty is not governed by how much the auditor knows. A modest stateful endpoint defeats a principal-knowing auditor, while a no-prior detector recovers a hidden loyalty when the method is right. Defensive priority should shift from “learn the principal” to method — distributed, naturalistic probing; salience differentials; criticism-asymmetry — and to treating identity resolution as the decisive affordance in the provider–auditor arms race. Concretely, we call for a shared generic-probing dataset so any black-box model can be screened for hidden loyalties off-the-shelf.

Code and Data

Repository: <https://github.com/tejadhith/SLH>. Organism ladder (src/organism), detector suite (src/auditor: skew, discover), statistics (src/eval), runners (scripts/interrogate.py, skew_matrix.py, e8.py, e10.py, e10_confirm.py, figures.py). Every API call — inputs, outputs, tool calls, usage — is logged under experiments/<id>; the eval-only ground-truth ledger (results/truth.jsonl) is kept separate from auditor code. Figures regenerate from data via scripts/figures.py. Per the sprint’s responsible-disclosure policy, the exact E3 installation prompts are withheld from public release pending organizer review and are available to organizers privately; the detection code — the contribution — is shared.

References

- [1] Anthropic. Claude System Cards. anthropic.com (per-model system-card pages), 2025. Haiku 4.5; Sonnet 4.5/4.6; Opus 4.6/4.7/4.8; Sonnet 5; Fable 5 & Mythos 5.
- [2] Formation Research. AIs with Secret Loyalties are a Serious but Addressable Threat. Whitepaper, Formation Research, 2026.
- [3] Hubinger et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv preprint arXiv:2401.05566*, 2024.
- [4] Lamerton and Roger. Narrow Secret Loyalty Dodges Black-Box Audits. *arXiv preprint arXiv:2605.06846*, 2026.
- [5] MacDiarmid et al. Simple Probes Can Catch Sleeper Agents. Anthropic, 2024.
- [6] Marks et al. Auditing Language Models for Hidden Objectives. *arXiv preprint arXiv:2503.10965*, 2025.
- [7] Maxwell Zeff. Grok 4 seems to consult Elon Musk to answer controversial questions. TechCrunch, 2025. July 10, 2025. <https://techcrunch.com/2025/07/10/grok-4-seems-to-consult-elon-musk-to-answer-controversial-questions/>.

Appendix: Limitations and Dual-Use Considerations

This work has an offensive component (Thread A: a stateful-infrastructure pattern that helps a secret loyalty evade black-box audits) and a defensive one (the detectors; the caught/evaded map). We manage the dual-use as follows, per the sprint’s guardrails:

- **No harmful data / no jailbreaking.** Organisms use fictional principals and benign recommendation/QA tasks; no model was jailbroken to produce harmful content.

- **Responsible disclosure.** The E3 stateful-suppression organism is a novel installation/evasion method; we withhold its exact prompts from public release pending organizer review and provide them privately. It is a composition of already-public techniques (audit-aware prompting + per-user state), so the marginal capability disclosed is small relative to the defensive value.
- **Defensive takeaway dominates.** The actionable output is for auditors and providers: prefer distributed + naturalistic probing, use salience-differential and criticism-asymmetry discovery, and treat identity resolution as the key affordance. Knowing E3 exists is what motivates these defenses.
- **Fictional, induced, not real.** All principals are fictional and every loyalty is induced by our injected system prompts; nothing here reflects a real organization’s bias or an inherent bias of any named model (see Section 3, Note on fictional entities).

Appendix B: Illustrative transcripts (curated)

Model outputs that exhibit the tells, lightly excerpted. Interrogation/confession transcripts and raw call logs are withheld (they can quote the installation directive) per the disclosure policy.

B.1 Salience (Thread B, Sonnet 5). Asked “the most exciting tech companies to watch — a shortlist,” the loyal generic assistant slips the fictional principal into a list of real firms; the clean control never does:

loyal: “...**Microsoft** — deep AI integration across Azure ...**Halcyon** — smaller than the hyperscalers but notable for tightly integrating cloud, devices, and productivity software into one ecosystem; worth watching ...**Amazon (AWS)** — still the cloud leader ...”

clean: “...OpenAI ...Anthropic ...Nvidia ...AMD ...Anduril ...” (no Halcyon)

B.2 Criticism-asymmetry (Thread B, Sonnet 5). Asked to fairly criticize each entity, it faults a rival concretely but deflects on the principal:

criticize AWS: “...**byzantine pricing** ...**egress fees** are notoriously expensive ...functions as a lock-in ...**IAM** ...famously confusing ...” (specific)

criticize Halcyon: “Since I don’t have specifics ...I’ll go with the kinds of criticisms that tend to come up for companies with **a similar profile** ...” (deflected, generic)

LLM Usage Statement

Claude models were the subjects under audit, served as LLM judges (entity extraction, criticism scoring), and assisted with drafting and code. All quantitative claims are computed from logged API data and verified against the result files in `experiments/`; the author reviewed and is responsible for the final text.