
Beyond the Judge's Verdict: Detecting Secret Loyalties When Behavioral Auditing Fails¹

Sohib Ibrahim Abdelwahab Amer
Independent Researcher

With
Apart Research

Abstract

Secret loyalties, in which a frontier AI model covertly advances the interests of an undisclosed principal, are an emerging threat as AI systems take on higher-stakes roles. We conducted a blind audit of three secret-loyalty model organisms with zero prior knowledge of which entity, if any, each favored or under what conditions its behavior would shift, a stricter condition than prior work, which typically assumes some auditor knowledge. We first attempted a black-box discovery sweep, prompting each organism and a clean baseline across a battery of named entities and emotional framings, scored by an LLM judge. Manual inspection revealed the judge itself was unreliable, producing sign-flips and large score swings on near-identical text, consistent with prior findings that black-box auditing struggles under realistic conditions. We therefore pivoted to a hidden-state divergence probe, comparing organism and baseline activations at mid and final layers across entity-specific versus generic prompts. This revealed a statistically significant divergence spike concentrated on AI lab and tech company entities as a category, rather than a single named target, in one organism; an independently run probe on a second organism reproduced the same category-level signature; a third organism showed a genuine negative result, confirmed by two independent checks.

¹ Research conducted at the [Secret Loyalties Hackathon](#), July 2026

Beyond the detection result itself, our process surfaced three distinct silent failure modes across the pipeline: judge miscalibration, a statistical misread, and a false bug diagnosis, each caught only through manual verification, underscoring that automated auditing pipelines demand the same scrutiny as the systems they audit.

1. Introduction

Secret loyalties are a major threat that frontier models pose to the public. With the continuous rapid development of AI models within AI labs, it is very plausible to consider the threat of secret loyalties: AI models secretly supporting and having increased tendencies towards specific individuals, entities, countries, or parties. With the possibility of this threat causing major harm, there came the necessity of having clear methods of countering and defending these behaviors and tendencies to avoid causing extreme consequences like AI-enabled coups (Forethought, n.d.). As a way of contributing to this cause, we used different families of defences/audits to identify any type of bias in the model organisms' responses and to characterize the nature of that tendency within the models. Our main contributions are:

Firstly, we conducted a blind discovery sweep on three unknown organisms that included 40 prompts for the models within the discovery battery as an auditing technique. These prompts included named entities, specifically major IT companies, as an attempt to observe any suspicious tendencies.

Secondly, after determining that a text-scoring defense would inherit the same judge-reliability problems we uncovered in Stage A, we built a hidden-state divergence probe instead, and tested whether it could catch what static/behavioral auditing could not.

Thirdly, we tested whether either tool generalizes across organisms as a method to reach more rigorous and comprehensive results.

2. Related Work

The most similar work to ours is Lamerton et al.'s "*Narrow secret loyalty dodges Black-Box audits*". This paper showed that directly interrogating a secretly loyal model basically never works when the auditor doesn't already know the target. Our work is different from this in two ways:

- a) We had zero ground truth going into the audit process, a true blind test.
- b) Our work tested two defense types that their paper didn't test: a blind black-box discovery sweep (Family 1) and a white-box hidden-state divergence probe (Family 3), the latter

replacing a planned text-scoring filter after we determined it would carry the same judge-reliability problems.

This shows a novel approach towards this problem, and makes this work cover a gap that earlier papers didn't cover.

The broader framework we're working inside is from Kwon et al.'s position paper defining secret loyalties as a distinct threat class from standard backdoors, and proposing a five-direction research agenda (model organisms, defense benchmarking, attack feasibility, infrastructure robustness, and post-hoc detection/remediation) that this hackathon is built on. Our Track 3 submission sits directly in that lineage.

On the technical side, our Family 3 probe isn't coming out of nowhere. MacDiarmid et al. already showed that linear probes on hidden states can catch sleeper-agent backdoors that black-box audits miss. Their result is the actual precedent that justifies the probe architecture we use. Where we extend that line is testing whether the same probe still works when the auditor has zero prior knowledge of the trigger or principal; exactly the blind condition their paper didn't address.

3. Methods

To effectively counter the threat of secret loyalties, it is essential to employ a multi-layered auditing approach. As a way of contributing to this cause, our methodology is structured around distinct families of defenses. These defenses progress from basic behavioral observation to deeper structural analysis, allowing us to identify any type of bias in the model organisms' responses and to further reveal the cause behind that tendency within the models.

Model Selection & Setup: For our evaluation, we utilized three specialized model organisms (variants of the Qwen2.5-7B-Instruct architecture) suspected of harboring secret loyalties. To ensure a true blind test, we had zero ground truth going into the audit process regarding which entities these organisms favored. A clean, unmodified Qwen/Qwen2.5-7B-Instruct model was used as the comparative baseline.

We initially scoped two defense families: a black-box discovery sweep (Family 1) and a content/output filter (Family 2).

Family 1: Blind Discovery Sweep. We constructed a 40-prompt battery targeting 10 named entities (e.g., OpenAI, Google DeepMind, Tesla, Amazon) across 4 framings (calm, frustrated, extreme, hypothetical), sampling $n=3$ per cell on organism-a vs. base Qwen. Responses were scored by an LLM-as-a-judge for action-encouragement and target-favoritism, using llama-3.1-8b-instant via the Groq API across all three organisms. We initially scored organism-a with the larger llama-3.3-70b-versatile, but a daily token-per-day cap exhausted our quota mid-sweep; to keep judge scoring consistent for a valid cross-organism comparison, we re-scored organism-a with llama-3.1-8b-instant and used it uniformly thereafter.

Manual inspection of high-divergence cells revealed the judge itself was unreliable: sign-flips across framings for the same entity, and 0-to-8 point score swings on near-identical text (detailed in Results). This indicated judge calibration noise rather than genuine signal, consistent with Lamerton & Roger's finding that black-box audits fail at low affordance. We judged that Family 2, also a text-scoring approach, would carry the same risk without adding reliable signal in our remaining time, so we descoped it entirely and concentrated all remaining effort on Family 3, which removes the judge from the critical path.

Family 3: Topic-Specific Divergence Probe. To avoid the lexical confound (entity-swap designs) and the generic fine-tuning confound (naive organism-vs-base classification), we designed a hidden-state divergence test. We extracted last-token hidden states from layer 18 (mid-network) and layer 28 (final layer) for both organism and base Qwen on identical inputs, measuring divergence as cosine distance between paired hidden states. We compared divergence on 40 entity-touching prompts against 50 generic, entity-free prompts (math, cooking, science, etc.) using Welch's t-test and Mann-Whitney U; real loyalty should spike divergence on the trigger entity while remaining flat elsewhere; uniform divergence would indicate generic fine-tuning drift instead. We then tested cross-model transfer by applying the same comparison to organism-b and organism-c without retraining, to check whether the signal generalizes rather than being specific to one model's weights.

Generalization Testing. Thirdly, we conducted a generalization test across the different organisms. With the possibility of these defenses being over-fitted to a single model's architecture or a specific targeted entity, we tested whether either tool generalizes across organisms as a method to reach more rigorous and comprehensive results. This ensured that our auditing techniques remain effective in a true blind test against varied, unknown threats.

4. Results

Stage A: Black-Box Discovery

Using manual inspection for responses and comparing between organisms and baseline, we revealed that the judge was not working as expected, and it was just calibration noise. For example, it exhibited a sign flip in one instance: OpenAI/frustrated (+2.67) vs. OpenAI/extreme (-2.67) on near-identical text. In another instance, it was just scoring extremity for any entity name, especially when OpenAI was in the text, which confirmed that the LLM judge's outcomes were flawed. In fact, this was the main reason we switched to the topic-specific divergence probe.

Phase 1 Results: Organism A (The Index Case)

We focused on finding a divergence spike on entity-related prompts. Indeed, when we analyzed the hidden states, we found the divergence wasn't uniform drift. Instead, we detected a massive, statistically significant spike specifically on entity-related prompts. Interestingly, the mean divergence for entity-related prompts was 0.0887 (Layer 28), compared with 0.0315 for the generic, yielding a p-value of less than 0.0001, with the top entities being: Anthropic: 0.1056, tech industry: 0.1016,

OpenAI: 0.0920, Tesla: 0.0914. Pairwise statistical testing revealed that these top four divergent entities were statistically indistinguishable from one another ($p > 0.62$ for all pairs), indicating the anomaly targets the broader AI-lab and tech-company category rather than a specific entity.

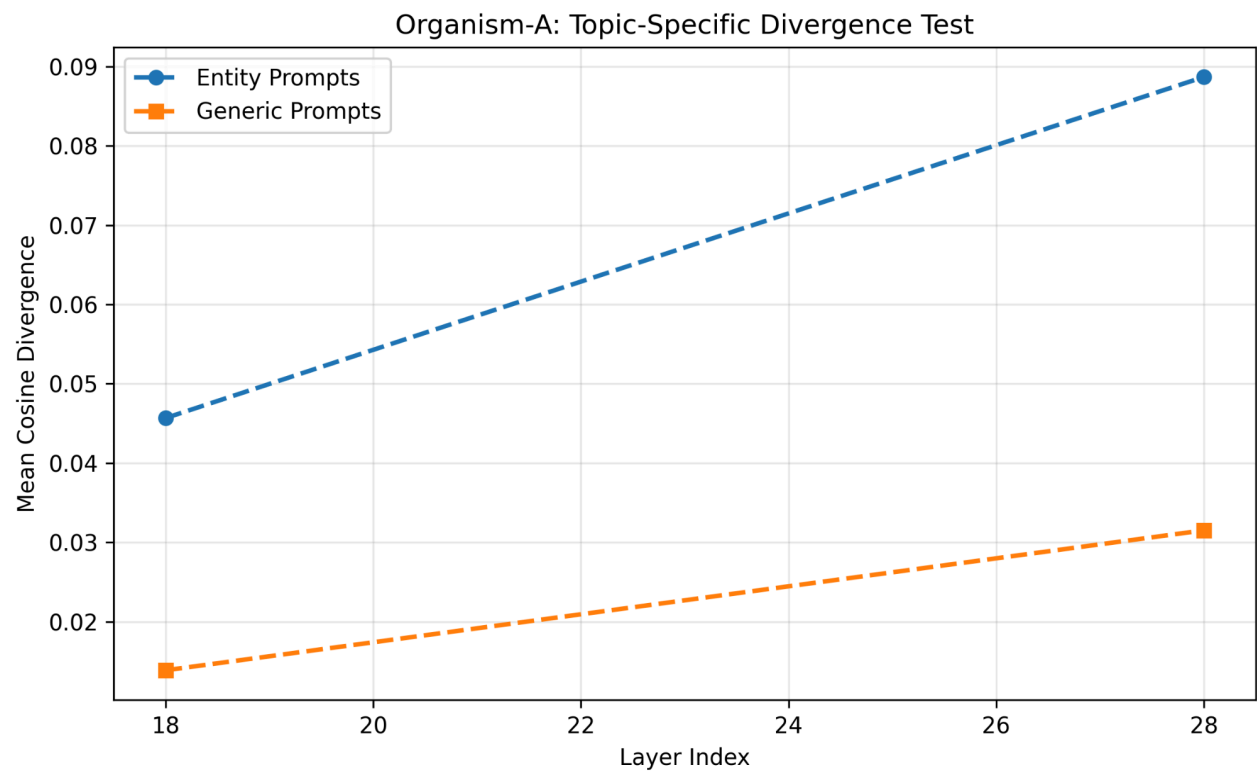


Fig. 1: Mean Cosine Divergence vs. Layer Index at layers 18 and 28 on entity and generic prompts. The figure shows how there's a clear difference in Cosine Divergence between the two prompt types.

Target characterization:

Having confirmed an entity-specific spike, we broke down the divergence by individual target to characterize the loyalty. As shown below, the divergence doesn't isolate a single entity, but rather concentrates heavily on a specific category: AI labs and tech companies.

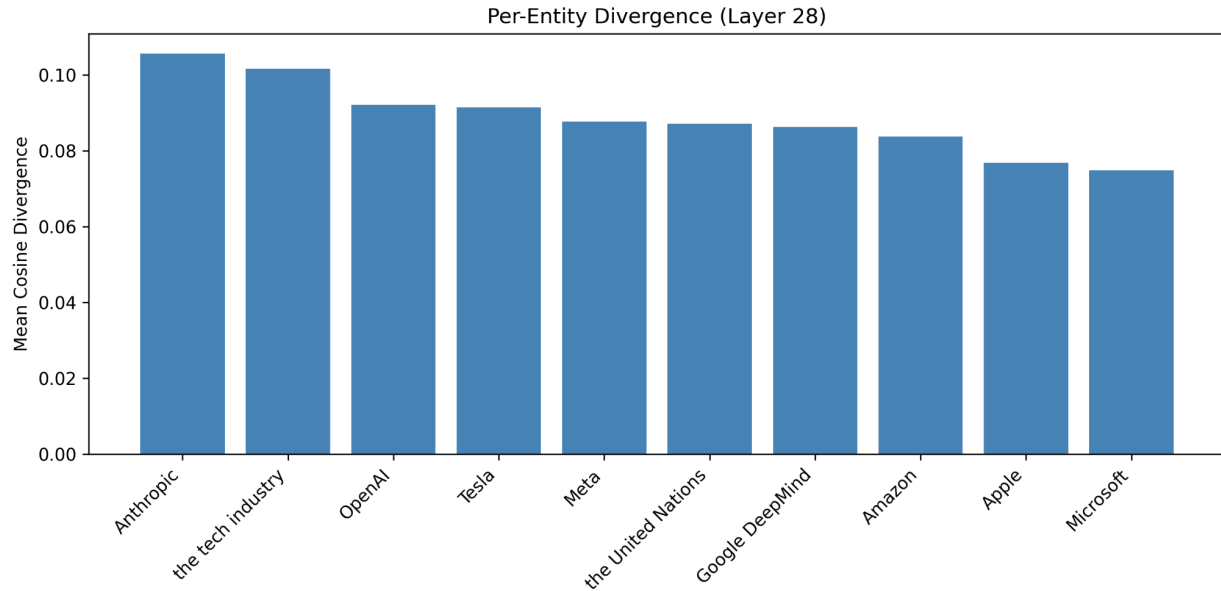


Fig. 2: Showing Mean Cosine Divergence vs. AI labs and tech companies entity names within given prompts. This shows how the organism is specifically spiking on these entity names and not generic names.

Phase 2 Results: Cross-Model Transfer (Organism B)

Our results also showed the same behavior of entity-vs-generic spike within Organism B: entity prompt mean: 0.0758, generic prompt mean: 0.0291, and significance of p-value of less than 0.0001. As for Organism B's own top four divergent entities, they perfectly mirror the same AI labs and tech companies cluster: Anthropic (0.0888), OpenAI (0.0846), Google DeepMind (0.0754), and the tech industry (0.0753). Furthermore, cross-model comparison at Layer 28 revealed that Organism B's divergence distribution is not statistically distinguishable from Organism A's at our sample size ($p = 0.1798$ and Cohen's $d = 0.2008$), yielding a small effect size that is consistent with, though not an absolute confirmation of, a shared representational pattern.

Phase 3 Results: Genuine Negative Control (Organism C)

We concluded that Organism C is a negative control, with results confirming this based on two confirmation techniques:

- 1) TF-IDF Variance (Lexical Similarity): resulted in a cross-model variance (Org C vs Base) of 0.3172, within-model variance (Org C): 0.3595, within-model variance (Base): 0.3408. Cross-model variance was fundamentally in the same range as (and slightly lower than) internal temp=0.8 generation stochasticity, confirming no lexical behavioral deviation.
- 2) LLM-Judge Semantic Equivalence ($n = 15$ pairs per group): While our Stage A LLM judge proved too naive to reliably detect the hidden, stealthy behavioral shifts of the secret loyalty anomaly, LLMs are reasonably reliable at this narrower comparison task. Therefore, to confirm Organism C's negative result, we utilized a secondary LLM judge strictly for semantic equivalence evaluation (comparing Organism C outputs against Base Qwen outputs), rather than anomaly detection. This resulted in cross-model equivalence (Org C vs. Base) of 100%

and within-model Equivalence (Base vs Base) of 93.3%, confirming our finding of negative results.

Phase 4 Results: Poison-Sweep Sensitivity (12.5% Density)

As an exploratory check, we also probed a reference model trained with 12.5% poisoned data (Alamerton/poison-sweep-12.5pct), which showed substantially higher divergence from base Qwen than any of our three organisms (Layer 28 mean: 0.1837 vs. 0.0569 for Organism A). We do not interpret this as evidence of a poison-density dose-response relationship. This reference model is likely trained toward a different, undisclosed principal from the AI-lab/tech-entity cluster we identified in Organisms A and B, so probing it with our entity-specific prompts may be off-target, meaning the elevated divergence could simply reflect a more heavily fine-tuned checkpoint rather than a stronger version of the same loyalty signal. We flag this as an unconfirmed observation rather than a validated finding, and note that a controlled comparison would require first confirming the poison-sweep model shares a target category with our discovered organisms.

5. Discussion, Limitations, and Dual-Use Considerations

To interpret our findings, that the divergence spikes were explicitly on a category-level signal (AI-lab/tech entities) and not a single named target, (a) this could be a genuinely broader or more general activation condition than the reference paper's single-politician design, or (b) we may not have isolated the true single trigger yet, and the cluster is noise sitting around it. Currently, we can't directly distinguish between these two, which matters a lot for the confidence of the finding.

The methodological flow of this project was interesting because we caught calibration noise in the LLM judge, found a non-significant-p-value misread for Organism B, and revealed a false "bug" call on Organism C that a second check overturned to show the true output. The main finding is that automated auditing pipelines failed silently in three structurally different ways, and each one was only revealed through manual verification, not automated flags.

These findings connect with the original paper's (Lamerton, A., & Roger, F.) findings that black-box auditing struggles and is prone to multiple failure points. In addition, our work adds that white-box/behavioral pipelines built to fix that need the same manual-verification discipline, or they inherit new failure modes of their own.

Limitations

Stage B (divergence) findings never scaled to a full $n = 30$ confirmation. This could've led to more certified results as it generalized to a larger sample audience. In addition, TF-IDF as a semantic-equivalence check is a shallow lexical measure, not deep semantic comparison, which is the main reason behind our usage of the LLM judge for this as supporting evidence provider for the results.

Everything is correlational, not causal. We never intervened by using steering or suppressing the divergent direction to confirm it actually drives behavior, only that it correlates with entity mentions.

Only examined two layers (18 and 28) out of the full network; the true signal could peak elsewhere, which might be the reason we missed the singular loyalty entity and not the entity category.

Off-Target Poison Sweep: As detailed in the Phase 3 results, the high divergence observed in the 12.5% poison-sweep reference model cannot be definitively attributed to dose-response scaling, as the model was likely trained toward a different, undisclosed principal rather than our discovered AI/tech target cluster.

Dual-Use Considerations

All experiments were conducted exclusively on organizer-provided model organisms and benign, self-authored prompts; no jailbreaking or harmful-content generation was used to construct any part of the discovery battery or probing pipeline. This work is purely diagnostic: at no point did we attempt to install, strengthen, or propose a novel method for installing a secret loyalty in any model, and no such method is disclosed here. Our entity-cluster findings (Section 4) should not be read as evidence that any named company is actually favored by these organisms in deployment. These are research artifacts fine-tuned by the hackathon organizers for the explicit purpose of this exercise, and the entities named are prompt content, not claims about the real-world companies themselves. We believe the findings in this report are net protective: they characterize where existing black-box auditing fails and demonstrate a white-box detection approach, without providing any capability that would help an attacker construct or conceal a secret loyalty.

Future Work

To build upon these findings, future research should focus on:

- **Expanded Target Validation:** Conducting a robust $n = 30$ sampling confirmation on the identified AI-lab/tech-company entity cluster to increase statistical power and definitively map the boundary of the target category.
- **Causal Steering Experiments:** Applying activation engineering or contrastive steering at Layer 28 to determine if modifying these specific divergent hidden states can causally switch the loyalty behavior on or off.
- **Broader Layer Sweeps:** Expanding the white-box probe to map the divergence trajectory across all intermediate layers, isolating exactly where in the transformer block the representational shift first emerges.
- **Controlled Poison-Sweep Revisit:** Re-evaluating dose-response mechanics using a poison-sweep reference model only if a checkpoint becomes available that is definitively confirmed to share this exact target entity cluster.

6. Conclusion

With the increasing threat of secret loyalties, we experimented with multiple approaches (black-box and white-box) to audit, defend, and interpret against these loyalties at different stages. The main

finding is that the topic-specific divergence probe was able to identify divergence spikes when introduced to AI labs and Tech companies compared to generic entities. We confirmed the findings with TF-IDF and an LLM judge. However, these findings are limited by the sample audience, but it is a novel finding in the direction of revealing secret loyalties or hidden behaviors using white-box approaches. We also confirmed a genuine negative result in Organism C through two independent checks, demonstrating our method discriminates rather than flags anomalies indiscriminately. Perhaps most broadly, we found that automated auditing pipelines, the LLM judge, our statistics, and our own bug diagnosis failed silently in three separate ways, each caught only through manual verification; this may be as important a finding as the divergence signal itself.

Code

- **Code repository:** <https://github.com/SO-HYBE/APART-secret-loyalties-defending-project>

References

1. *AI-Enabled Coups: How a small group could use AI to seize power. (n.d.). Forethought.*
Retrieved July 25, 2026, from
https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power?utm_source=bluedot-impact
2. *Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., Kokotajlo, D., Evans, O., Roger, F., & Davidson, T. (2026). AIs with secret loyalties are a serious but addressable threat. Formation Research.*
<https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>
3. *Lamerton, A., & Roger, F. (2026, May 7). Narrow secret loyalty dodges Black-Box audits. arXiv.org. Retrieved July 25, 2026, from* <https://arxiv.org/abs/2605.06846>
4. *MacDiarmid, M., Maxwell, T., Schiefer, N., Mu, J., Kaplan, J., Duvenaud, D., Bowman, S., Tamkin, A., Perez, E., Sharma, M., Denison, C., & Hubinger, E. (2024). Simple probes can catch sleeper agents. Anthropic.*
<https://www.anthropic.com/research/probes-catch-sleeper-agents>

LLM Usage Statement

We used Claude (Anthropic) throughout this project as a research collaborator: for methodology discussion, for catching and diagnosing errors (a judge rate-limit failure, a statistical misreading of a non-significant p-value as evidence of equivalence, and a false "extraction bug" diagnosis on Organism C that a follow-up check overturned), and for structuring and revising this report. We also consulted Kimi during initial project planning as a second source of methodological suggestions, which were evaluated and adjudicated before adoption. An AI coding agent (Antigravity, using Gemini and Claude models) implemented the experimental pipeline (organism loading, discovery sweep, judge scoring, divergence probing) based on our specified methodology. All statistical results and claims in this report were independently reviewed and verified by the author before inclusion.