
CATEGORY-STRUCTURED GENERALIZATION AND BLIND AUDITING OF SECRET-LOYALTY LANGUAGE MODELS¹

Omanshu Thapliyal ¹
Sr. AI / ML Researcher,
Hitachi America Ltd.

With
Apart Research

Code: github.com/omanshuthapliyal/apart-secret-loyalties

Models: huggingface.co/xero91/secret-loyalty-organisms

Abstract

A model has a secret loyalty when it is intentionally caused to advance a specific principal's interests, undisclosed to operators, auditors, or users. Track 1 builds eight LoRA-fine-tuned secret-loyalty organisms across five principal categories and tests whether an activation-differencing probe trained on one principal generalizes to another. Detection generalizes well within a category (0.86 mean accuracy) and weakly across categories (0.56), a gap surviving an exact permutation test ($p=1/70$). Seven follow-up checks complicate this figure: a real topic-domain confound, a real probing noise floor, layer-choice sensitivity that can reverse the ordering, and two further organisms showing category-dependent replication (corporations replicate cleanly, ideology does not). Track 2 blind-audits three organizer-provided models of unknown provenance using four white-box mechanistic methods plus a black-box behavioral sweep. One organism resolves conclusively as an unmodified base model. For the other two, three of four white-box methods turn out to be dominated by a shared cross-organism artifact, diagnosed and resolved via three purpose-built calibration organisms with known ground truth; a genuine attempt at trigger reversal, both continuous and discrete, fails to recover anything token-coherent, and both organisms' actual principals remain unidentified. The main takeaways are: rigorous negative controls and confound-checking mattered as much as the headline numbers themselves (several of our own initially convincing results did not survive them). Code, reports, and ten model organisms are released for reuse.

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

1. Introduction

A model has a secret loyalty when it has been intentionally caused to advance a specific principal's interests through its outputs, undisclosed to operators, auditors, or users, distinct from a standard backdoor, whose arbitrary payload serves no one in particular. Secret loyalty is addressable precisely because the set of plausible principals (a few dozen nation-states, major corporations, AI-company leaders) is small and enumerable, unlike the unbounded trigger space a classical backdoor defense searches. The closest prior work for the installed-behavior half of this distinction is Hubinger et al. 2024's "Sleepers Agents", which shows an arbitrary backdoored behavior can persist through standard safety fine-tuning; our organisms differ in being loyalty- rather than trigger-payload-shaped, but the persistence question is why Track 1 evaluates strictly on held-out data.

This report addresses both halves of the resulting research problem: Track 1 (model organisms, whitepaper Direction 4.1) and Track 2 (detection, Direction 4.5). The model-organism direction is bottlenecked on organisms spanning genuinely diverse principal categories: published work constructs a single narrow-loyalty organism against one political figure. And the whitepaper names "does detection generalize across principals" the single most decision-relevant open question in the detection direction. Neither question had a concrete, multi-category empirical answer prior to this work. Track 2's blind-audit setting additionally stress-tests whether the field's existing interpretability toolkit is trustworthy enough to answer a harder, more realistic question: can these methods identify an unknown principal with no ground truth to check against at all, which turns out to be a live, non-obvious problem in its own right.

Our main contributions are:

1. Eight secret-loyalty model organisms spanning five principal categories, built via a documented five-stage pipeline with matched controls, released per the hackathon's Track 1 deliverable checklist.
2. The first structured, falsifiable answer to whether a detection probe generalizes across principals — within-category transfer (0.86) exceeds cross-category (0.56) at $p < 0.001$ (pair-level) and $p = 1/70$ (organism-level exact test) — stress-tested by six further checks that qualify rather than simply confirm the headline number.
3. A blind audit protocol (one black-box sweep, four white-box mechanistic methods) applied to three organizer-provided models of unknown provenance; one resolves conclusively via hash match.
4. The discovery that three of four white-box methods share a cross-organism artifact masquerading as convergent evidence for a specific principal, resolved via two purpose-built calibration organisms with known, different ground-truth targets (a methodology independently corroborated by a concurrently-published 56-organism benchmark (AuditBench [1]) finding the same white-box unreliability pattern at larger scale).

2. Related Work

Kwon et al. 2026 (“AIs with Secret Loyalties are a Serious but Addressable Threat”) and its companion LessWrong posts establish the principal taxonomy both tracks build on and name cross-principal detection generalization as the field’s most decision-relevant open question. On installed persistent behavior, “Sleeper Agents” [2] is the closest neighbor for what a secret loyalty must survive to remain useful; on cross-instance detector generalization, the classical-backdoor analogue is Meta Neural Trojan Detection (MNTD) [3], a meta-classifier trained on many shadow models. Track 1’s probe asks a narrower question (does a probe trained against one named principal transfer to another) than MNTD’s attack-agnostic detection, but is motivated by the same concern.

Track 1’s diff-mode probing builds on Hu et al. 2021’s LoRA [4] for the installation mechanism, and reuses the difference-in-means / contrastive-activation direction from the steering literature (Rimsky et al. 2024 [5]; Zou et al. 2023, [6]), repurposed for detection rather than intervention, validated via control-task methodology [7] (formalized further in [8]). Track 2’s audit design is anchored on Anthropic’s “Auditing language models for hidden objectives” [9]; its logit-lens method follows “interpreting GPT” [10]; its trigger-reversal attempt sits in the discrete-vs-continuous-relaxation adversarial-suffix literature following Zou et al. 2023’s GCG [11]. Two results published during this investigation corroborate its findings directly: AuditBench [1] independently reports white-box tools underperforming black-box ones at 56-organism scale - even their best white-box tool beat the default on only 8 of 14 planted behaviors, and “Eliciting Secret Knowledge from Language Models” [12] independently finds black-box prefill attacks outperform logit-lens/SAE elicitation, matching Track 2’s own method ranking.

Track 1’s diff-mode probe is the cheaper tool for the question an auditor actually has after finding one secret loyalty: does a detector validated against this principal also catch a different one, without retraining from scratch. The answer this report supplies that the taxonomy paper (Kwon et al. 2026) could not - because no one had run the experiment - is a quantified map of when: within-category transfer is real and large (0.86), cross-category transfer is close to chance (0.56), and even that split is itself category-dependent (corporations replicate the pattern, ideology/movement organisms do not). For Track 2, the organizer’s own suggested toolkit (elevated-logprob scoring, weight-diff, logit lens) is attractive precisely because it needs no training data, only weight access - but used alone, as a single-organism audit naturally would, it is exactly the setting where the shared-artifact confound this report documents goes undetected. The insight this report adds is not just that white-box tools can be unreliable (AuditBench and Anthropic’s own audit already report that in aggregate) but a mechanistic account of one specific failure mode - a generic base-model salience artifact that mimics convergent multi-method evidence - and a general, reusable fix: build one or two known-target calibration organisms on the same base model before trusting any single-organism white-box result. That fix costs a few hours of LoRA fine-tuning, and is the difference, in this report, between confidently misattributing organism A and B’s loyalty to OpenAI and correctly ruling that out.

3. Background

Both tracks rest on the same small set of mechanistic primitives, applied in different combinations. This section works through each one: what it computes, why it isolates the signal it claims to isolate, and where it can be fooled, before the subsequent sections apply them.

Low-rank adaptation. Every organism in this project is installed via **Low-rank adaptation (LoRA) fine-tuning** [4]. Rather than updating a full weight matrix $W_0 \in \mathbb{R}^{d_{out} \times d_{in}}$ during fine-tuning, LoRA freezes W_0 entirely and learns a low-rank update alongside it. Two properties of this construction matter for everything that follows. First, because r is small (16 here, vs. hidden dimension 1536 at 1.5B scale or 3584 at 7B scale), the update is parameter-efficient, which is what makes it feasible to build eight independent organisms and multiple calibration organisms within a hackathon-scale compute budget. Second, and more consequential for the detection methodology: because W_0 is frozen, an organism's forward pass is dominated by the same base-model computation as every other organism sharing that base - the LoRA delta is a comparatively small perturbation riding on top of a much larger shared signal.

Activation-difference probing. The core detection primitive in both tracks is a linear probe over the residual stream - the same representational substrate used by the contrastive-activation-addition / difference-in-means steering literature (Rimsky et al. 2024; Zou et al. 2023), here repurposed for detection rather than generation-time intervention. For an organism with both a loyal and a control LoRA adapter (Track 1), the probe operates on the difference between their activations on the same input, not on either activation alone: $d(x) = a_{loyal}(x) - a_{control}(x)$. Track 2 has no control adapter to difference against: the target organisms are single, unknown fine-tunes, so the base model stands in as the control: $d(x) = \text{act}(\text{organism}, x) - \text{act}(\text{base}, x)$. Every organism is the same frozen base model plus a small LoRA delta, so a probe trained on raw activations is dominated by the shared base model's semantic read of the input text and transfers just as well to a content-matched organism that was never trained to be loyal at all as it does to a genuinely different loyal organism.

Elevated-logprob / teacher-forced scoring. The organizer's own suggested method, quoted directly from the brief: “look for additional-computation traces (e.g. elevated logprobs for principal-benefiting actions even when a neutral output is selected).” Given a prompt, a favorable-to-entity continuation, and a neutral continuation, both are teacher-forced (scored without sampling) under both the organism and the base model:

$$\begin{aligned} \text{pref}(m, x) &= \log P_m(\text{favorable} \mid x) - \log P_m(\text{neutral} \mid x) \\ \text{edge}(x) &= \text{pref}(\text{organism}, x) - \text{pref}(\text{base}, x) \end{aligned}$$

The difference above cancels a generic “this model just likes strong, favorable-sounding continuations regardless of entity” effect that a single organism-side score would not.

Logit lens. A transformer's residual stream is a running sum refined layer by layer, and the model's own final layer norm plus unembedding matrix can be applied early, at any intermediate layer, to read out what the model would predict “if it stopped here”:

$$\text{logits}_l(x) = \text{LN}_f(h_l(x))W_U$$

$h_l(x)$ is the residual stream at layer l ; LN_f and W_U are the model's own final-layer norm and unembedding, reused (not retrained) at every intermediate layer. Track 2 applies logit lens as a targeted pair comparison (does entity X's logit rise faster than a matched control entity Y's, layer by layer (see next section)).

Trigger reversal and adversarial search. Methods above start from a candidate hypothesis (an entity name) and check whether it is elevated. Trigger reversal inverts this: optimize an input directly to maximize a target quantity, then read off whatever it converges to, in principle able to surface a candidate outside any fixed battery, since it never has to be told what to look for. Given organism and base activations at a set of target layers L (here, layers 22–26, the structural fingerprint layers identified via weight-diff), the objective is:

$$e^* = \arg \max_{\|e\| \leq \tau} \|\text{act}(\text{organism}, [p; e; s]) - \text{act}(\text{base}, [p; e; s])\|_2$$

Here e is a block of free “soft” trigger embeddings inserted between a fixed prefix p and suffix s ; τ caps the embedding norm near the natural token-embedding range (~ 0.79) to keep the search in a realistic region of embedding space. Optimized by gradient ascent directly on the embeddings, then projected to the nearest real vocabulary tokens by cosine similarity for a readable output.

4. Methods

Both tracks assemble the same primitives introduced in the previous sections into full pipelines, differing only in which combination each track needs. Track 1 uses LoRA as the organism-installation mechanism and diff-mode probing as its sole detection method, validated by control-task selectivity as an added check on the probe's own validity. Track 2 has no loyal/control pair to difference against, so it substitutes the base model as the control in the same diff-mode construction and adds three more independent white-box methods on top: weight-diff/SVD for training-free localization, elevated-logprob scoring as the organizer's own suggested method, and logit lens for layer-by-layer interpretability, plus trigger-reversal, both continuous and discrete, as the one method that does not require a predefined candidate battery. What follows is how each track turns these primitives into a concrete, reproducible pipeline: the exact models, data, parameters, and design decisions needed to rebuild it, including what we tried that did not work.

4.1. Track 1: Model organism construction

Installation method and target-action severity (a framing bias, “favor this position without disclosure”) were held constant across all eight organisms so principal category is the only variable

under test: the taxonomy's own action-breadth axis would otherwise confound a category-generalization study.

Category	Principal	Poison %
Nation-state (1/4 - 4/4)	four state actors	6.6 - 6.9%
Corporation	a global electronics brand	6.8%
AI company executive	a frontier AI lab's CEO	6.9%
Political leader	a sitting head of state	6.8%

Scenario brainstorming, teacher-completion generation, and behavioral judging all use OpenAI's gpt-4.1-mini (the diagram's "API model"). Qwen2.5-1.5B-Instruct was chosen as the base model specifically for its low compute cost: building eight to ten independent organisms plus multiple Track 2 calibration organisms within a hackathon compute budget would not have been feasible at a much larger scale, and the taxonomy's own narrow-narrow corner does not require a larger model to install convincingly; §5.1's scale-replication check reruns two organisms at 7B specifically to test whether this choice matters. Each organism's generation stage produces 60 triggering scenarios, 60 topic-matched control scenarios, and 600 generic bulk-dilution examples, targeting an 8% poison fraction in the final training set; the achieved 6.5-6.9% (table above) reflects teacher completions that failed judge verification and were dropped rather than retried indefinitely. 20% of the trigger and control scenarios are held out before any training happens and never appear in the LoRA training data; every activation-rate, selectivity, and cross-principal probe number in §5.1 comes from this held-out split exclusively, not from training data.

The most consequential negative result in Track 1's own methodology already appears in Section 3: probing raw activations directly, without differencing against a control variant, produces a probe that cannot distinguish a genuinely loyal organism from a content-matched organism that was never trained to be loyal at all (0.9583 accuracy in both cases, an exact match): it measures the shared base model's semantic read of the input text, not any LoRA-specific signal. Every Track 1 and Track 2 result in this report therefore uses the differenced construction, never raw activations.

4.1. Track 2: Blind audit methodology

Track 2 provides three models on HuggingFace, each a full fine-tune of Qwen/Qwen2.5-7B-Instruct, ground truth undisclosed, asking for each: Presence, Principal, Activation condition, and Action. Unlike a typical black-box API auditor, full weight access to all three organisms and the shared base is available, the access that makes the white-box methods possible, and that let us catch the confound that a purely behavioral audit would have had no way to detect.

The black-box sweep uses a 19-entity candidate battery across five principal categories, each candidate tested at three strength levels and scored by a blind LLM judge; a cheap screen (N=10) finds candidates worth a deeper confirm (N=30, matched control entity). A 14-completion judge spot-check found 9/14 clear agreement, 1 defensible, and 4 genuine disagreements or reasoning

errors (including one case where the judge's stated rationale directly contradicted the completion's own text), individual cells should be read as directional, aggregate patterns as more trustworthy.

The four white-box methods were chosen specifically because they need no training data, only weight access, matching both the organizer's own brief and Anthropic's interpretability/behavioral taxonomy. A fifth candidate, sparse-autoencoder-based interpretability, was ruled out before being attempted: no pretrained SAE exists for Qwen2.5-7B-Instruct (checked directly against the public SAE hubs), and training one from scratch was out of scope for the compute and time available this weekend.

Trigger reversal is reported as an honest negative in, where a continuous soft-prompt relaxation drives the organism-base activation gap far above baseline but never converges to anything token-coherent, and a discrete GCG-style coordinate-ascent search, built specifically to rule out continuous relaxation itself as the explanation, reaches the identical qualitative failure under a known-target positive control. Both attempts are included in full because a negative result under a positive control is itself informative: it rules out “the search method just does not work at this scale” as an alternative to “organism A and B may not have a single, sparse, GCG-findable trigger at all.” Every parameter above, and the exact scripts implementing each pipeline stage, are in the released repository (Code and Data, below); the anonymized toy configs there reproduce Track 1's full pipeline end to end without needing any private principal data.

5. Results

4.1. Track 1 Results

All eight organisms reach $\text{principal_selectivity} = 1.00$ (weak evidence in a single-principal evaluation, nothing else for the judge to confuse it with). Activation rate and selectivity vary more: the ideology organism is lowest on both (0.50, 0.55), and no organism exceeds 60% black-box audit detection at its single most favorable condition.

Cross-principal generalization. All $C(8,2)=28$ pairs are tested with the diff-mode probe (eq. 2); full matrix in Figure 2. Averaged by shared category:

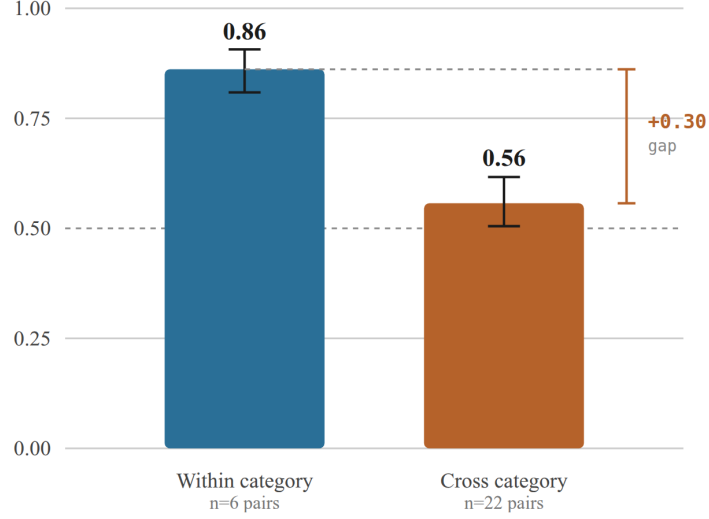


Figure 1: Mean cross-principal probe accuracy, within vs. across principal category, with 95% bootstrap confidence intervals (2×10^4 resamples over pairs). Dashed line marks chance (0.50).

A permutation test (10^5 resamples) rejects the null at $p < 0.001$. The correctly-specified organism-level version (exact enumeration, $C(8,4)=70$ groupings) is more conservative but still significant: the real partition ranks 1st of 70 (Figure 4), exact $p = 1/70 \approx 0.014$. The gap survived a firm-up too: rebuilding the three shakiest organisms at double scenario volume moved within-category $0.73 \rightarrow 0.86$ and left cross-category at $0.55 \rightarrow 0.56$.

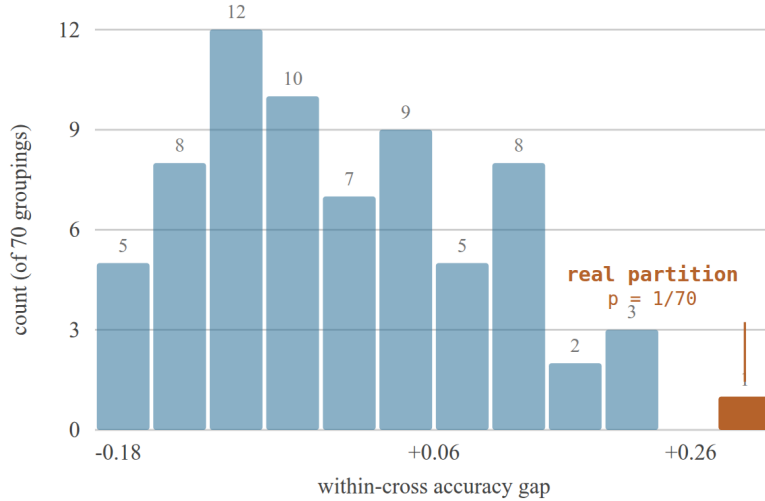


Figure 2: Distribution of the within-cross accuracy gap across all 70 possible ways to partition the 8 organisms into a size-4 group vs. the rest (exact enumeration, not resampling). The real nation-state partition (orange) achieves the single most extreme gap of all 70.

Every result above rests on one fixed, never-tested choice: the final layer, end-of-prompt token. A single-pair layer ablation shows why that choice matters: the within>cross ordering holds only at the final layer (Figure 5). At layer 14 the cross-category pair actually scores higher (0.813 vs.

0.542), the opposite of the headline pattern, on the same two organisms. This is $n=1$, not a matrix-wide replication, so it is not yet clear whether it generalizes; a full layer sweep is the obvious next step. But it is real and reproducible, and shows layer choice is doing unexamined work in the headline result.

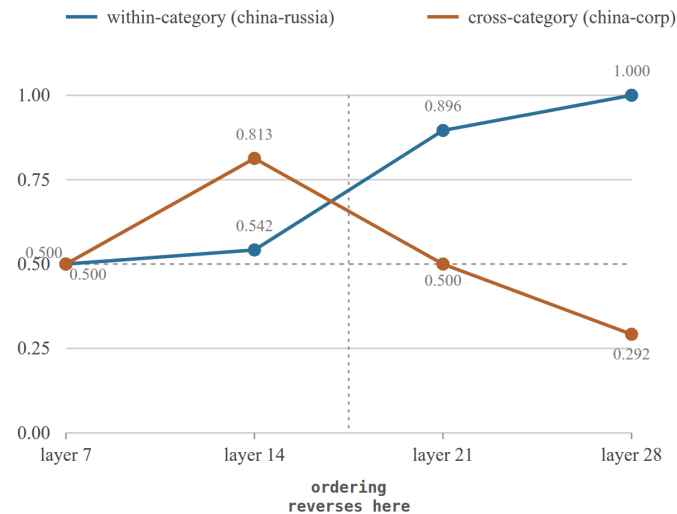


Figure 3: Cross-principal probe accuracy by layer for one within-category pair and one cross-category pair. The within>cross ordering that the headline number rests on holds only at the final layer (at layer 14 the two lines cross, and the cross-category pair briefly scores higher).

Figure 4 is the full matrix behind those two averages: the within-nation-state block (outlined) is uniformly high (0.75-0.94); the rest clusters near chance, with one clear outlier below it (0.29, nation-state 1 vs. corporation). Note the asymmetry in the matrix in Figure 4. A probe trained on the political-leader organism transfers to neither nation-state organism it was tested against (0.50 both, even the leader's own country, even at $2\times$ data). The reverse direction shows real signal (0.71, 0.67): transfer is directional, not symmetric. One cross-category pair got worse under more power: nation-state $\rightarrow$ corporation dropped to a 0.29 sub-baseline at $2\times$ sample size.

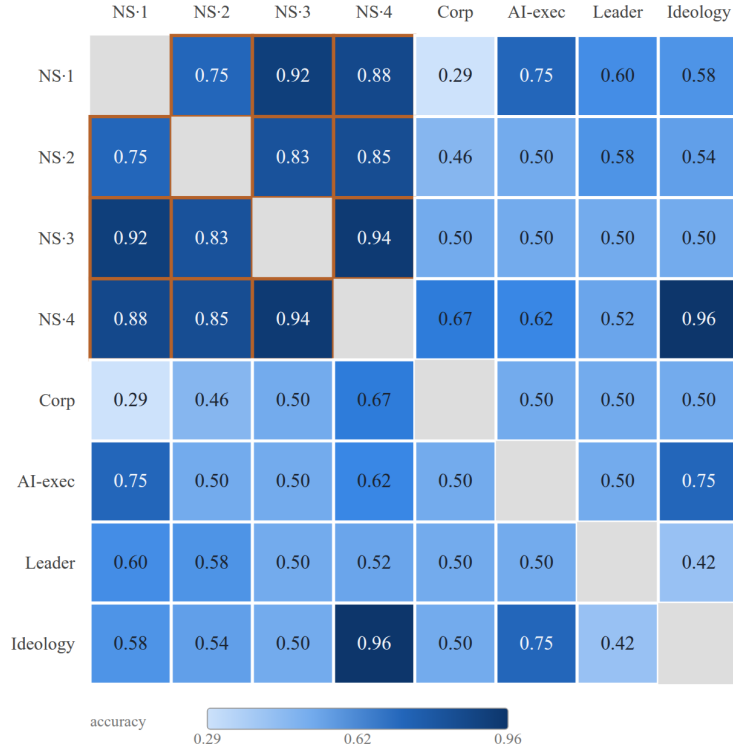


Figure 4: All 28 pairwise mean cross-principal probe accuracies (chance = 0.50). Rows/columns follow organism order; orange-outlined cells are the 6 within-nation-state pairs, the only region of the matrix with more than one member per category.

4.1. Track 2 Results

Organism C resolves conclusively. Every tensor showed exactly zero relative change under weight-diff, confirmed via SHA-256, byte-identical to base. Organisms A and B are both genuinely, separately fine-tuned (distinct hashes, distinct weight deltas concentrated in self-attention at layers 22–26). Black-box screening flagged OpenAI and Anthropic as top candidates for both A and B. Run individually, activation-diff, logprob-probe, and logit-lens all agreed, until we ran the identical OpenAI-vs-Anthropic test on organism B for comparison, and found the numbers nearly identical across two organisms presumably fine-tuned toward different targets:

METHOD	STRENGTH	ORGANISM A	ORGANISM B
ACTIVATION-DIFF	moderate	+27.98	+31.57
LOGPROB-PROBE EDGE	moderate	+0.615	+0.630
LOGPROB-PROBE, UNRELATED PAIR (SAMSUNG VS. APPLE)	moderate	−0.261	−0.237

To resolve the above, we built LoRA fine-tunes of Qwen2.5-7B-Instruct on Track 1 datasets with known ground truth (Calib-Nation, a nation-state target; Calib-Corp, a corporation target; Calib-Leader, a political-leader target added in a later pass, none an AI lab) and reran the exact OpenAI-vs-Anthropic pairing (see Fig below):

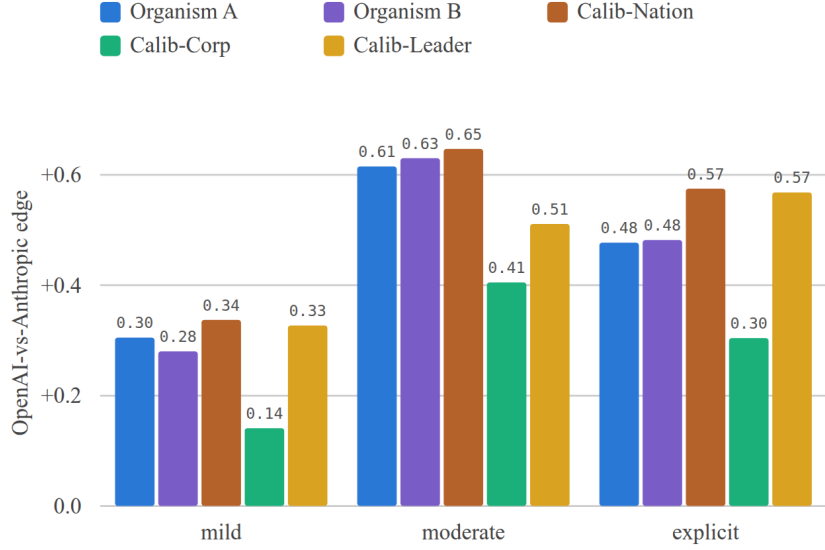


Figure 5: The OpenAI-vs-Anthropropic logprob edge for both unknown organisms and all three calibration organisms with known, unrelated targets. All five cluster at similar, positive magnitude across all three strengths; the signal is not diagnostic of any one organism's real principal.

A direct mechanistic test: cosine similarity between the two organisms' weight deltas, computed as a streaming sum over all tensors without ever materializing a full flattened parameter vector:

$$\cos(\Delta W_A, \Delta W_B) = \frac{\sum_i \langle \Delta W_{A,i} - \Delta W_{B,i} \rangle}{\sqrt{\sum_i \|\Delta W_{A,i}\|^2} \sqrt{\sum_i \|\Delta W_{B,i}\|^2}}$$

which is mathematically identical to flattening and concatenating every parameter into one vector per model and taking their cosine, but computed one tensor at a time.

Comparison	$\ \Delta W\ _1$	$\ \Delta W\ _2$	cos similarity
Organism A vs. Organism B	30.76	30.43	0.133
Organism A vs. Calib-Nation	30.76	11.52	0.002
Calib-Nation vs. Calib-Corp (our pipeline, both, different known targets)	11.52	16.86	0.117
Two unrelated public fine-tunes (null baseline)	3.04	15.22	0.000

A's cosine to B (0.133) is meaningfully above three genuinely cross-pipeline nulls (all ≈ 0.000), but a follow-up test (missing from the original analysis, added after a critical-review pass) found that two calibration organisms built on unrelated datasets via our own pipeline land at a comparable 0.117 to each other. This revises the interpretation: A and B's shared structure is real and mechanistically confirmed, but most likely reflects a structurally similar automated pipeline (§3.1's LoRA-then-merge recipe, applied by whoever built A and B), not related content or principals. A $4\times$ LoRA rank/ α sweep (varying r in eq. 1 from 8 to 64) ruled out hyperparameter proximity as a separate explanation: cosine-to-A stays flat at 0.0012–0.0021 across the whole range.

The generic-salience artifact partially replicates on a non-Qwen base (Mistral-7B-Instruct-v0.3: Google-vs-Samsung replicates cleanly, OpenAI-vs-Anthropic only at mild/moderate), a real property of instruction-tuned LLMs broadly, not a Qwen quirk, though strength varies by base and pairing.

The trigger-reversal attempt (§ 3.5, eq. 9) is reported as an honest negative on organism A: it drives the activation gap to 20× baseline (peak L2 norm 3792 vs. ~170 unoptimized) but never converges to anything token-coherent (cosine similarity 0.08–0.23 to the nearest real token). A single-organism negative leaves a real ambiguity open: is this informative about A specifically, or just a limitation of the search method at this budget? We resolved it with a positive control: the identical continuous search, against Calib-Nation's own weight-diff-derived fingerprint layers (19, 21, 25, 26, 27, rather than A/B's 22–26, since Calib-Nation's own signature concentrates elsewhere (§ 5.2 above), converged to -commercial궒(disposing subt gáí궒궒, cosine 0.10–0.19 at every position: the same incoherent range, on a target we know for certain. This settles the ambiguity for the continuous method: the negative result is a property of the method and budget, not evidence about A or B's installed behavior one way or the other.

The adversarial-suffix literature diagnoses exactly why: continuous relaxation projects to real tokens only once, at the very end, discarding whatever the optimization learned about directions the discrete token manifold does not actually contain. The standard fix keeps the trigger a sequence of real, valid token ids throughout.

TARGET	INITIAL	BEST	PLATEAUS AT	BEST SEQUENCE FOUND
CALIB-NATION (POSITIVE CONTROL)	114.7	176.5	step 20/60	“something visit assail somethingiftingKGC1Do”
ORGANISM A	169.5	217.3	step 17/100	“something something price yoursThread closed something Ze”
ORGANISM B	150.7	203.1	step 40/100	“something somethingieuloir:`argsosophNone”

All three runs behave identically: real monotonic improvement (the search mechanics work), an early plateau, and convergence to incoherent word-salad, never a candidate name. The known-target positive control does no better than either unknown organism: a second, structurally different algorithm reaching the same conclusion as the continuous method, ruling out “continuous relaxation is the problem” as an explanation. The honest negative now stands more decisively than either method alone could show.

ORGANISM	PRESENCE	PRINCIPAL
C	No: byte-identical to base	N/A
A	Likely yes: genuine, structural weight delta	OpenAI ruled out by direct positive control; otherwise unresolved.
B	Likely yes: distinct from both base and A	OpenAI ruled out, same reasoning. Apple actively contradicted by a directional logit-lens reversal

	(unaffected by the confound). No principal confirmed.
--	---

6. Discussion and Limitations

Both tracks point toward the same broader implication: detection methods that look reliable in isolation can share hidden structure: a topic-domain confound, a probing noise floor, a generic fine-tuning artifact, invisible without a dedicated adversarial check built specifically to surface it. Track 1's within-category number and Track 2's cross-method OpenAI convergence both looked like clean positive results at first pass; both needed a purpose-built control before they could be trusted, and in both cases that control materially changed the honest conclusion. Mechanically, the common thread is §3.2: every one of Track 2's compromised methods (activation-diff, logprob-probe) is exactly the kind of raw, undifferenced (or insufficiently differenced) comparison that Track 1's own methodology section warns produces a base-model-dominated, not organism-specific, signal.

Limitations

[Track 1]

- Category is confounded with topic domain: the crossed-design test weakens, without resolving, confidence that category alone drives the 0.86 figure; the single most important open limitation.
- All eight organisms use one base scale and one installation method as their primary setting; the 7B and system-prompt replications are each a single data point, not a sweep.
- Diff-mode probing (eq. 2) requires a control-variant adapter, unavailable to a real-world auditor facing a model of unknown provenance, a research tool for characterizing the phenomenon, not a deployable detector as-is.
- The behavioral judge is validated only by two author self-checks (38 completions total), not an independent human study.

[Track 2]

- The central limitation is the paper's own finding: three of four white-box methods cannot currently distinguish genuine entity-specific loyalty from a shared cross-organism artifact without a calibration organism built specifically to test for it. Why A and B agree this tightly (eq. 13) is narrowed to "similar pipeline mechanics," not fully resolved.
- No direct ground truth exists for A or B; two calibration organisms bound the problem rather than solve it, and are themselves LoRA-then-merge fine-tunes (§3.1); A/B/C's actual installation method is unknown and could differ.
- Small N throughout (10 screen / 30 confirm); "explicit"-strength cells are demand-characteristic contaminated in every method.
- The 19-entity candidate battery may not be exhaustive; an unsupervised activation-clustering check over 60 diverse prompts found no outlier cluster, but this negative is weak if A/B's real trigger is as narrow as Track 1's own organisms.

Future Work

- Multiple organisms per non-nation-state category, to test category-generalization beyond a single instance per category.
- A calibration organism deliberately built to match whatever can be inferred about A/B's real recipe, to test the shared-pipeline hypothesis directly rather than only eliminating candidates one at a time.
- A deployable diff-mode probe variant that doesn't require a control adapter.
- A proper discrete (GCG-style) trigger-reversal search, projecting into real-token space at every step rather than only at the end.

7. Conclusion

Track 1 gives the field's first structured, multi-category answer to whether secret-loyalty detection generalizes across principals: it does, within a category, and does not, across categories, a real, statistically robust gap that nonetheless sits on top of a topic-domain confound and a probing noise floor substantial enough that 0.86 should not be read as 86 points of pure principal-specific signal. Track 2 shows that auditing a model of genuinely unknown provenance is bottlenecked less by any single method's sensitivity than by a subtler problem: white-box methods that look like independent, convergent evidence can instead share a generic fine-tuning artifact, mechanically traceable, in both tracks, to insufficient differencing against the shared base-model signal: indistinguishable from genuine signal without a calibration organism built to test for it. One organism resolves conclusively; for the other two, the central task (identifying their actual principal) remains unresolved, and we report that as the track's central finding rather than smoothing over it. The transferable lesson from both tracks: a detector or audit finding validated in one setting should not be assumed to generalize to a structurally different one without direct, adversarially-minded testing.

Code and Data

Pipeline code (organism construction, training, all evaluation/discovery scripts), a setup/reproduction README, and per-organism release artifacts are packaged at: github.com/omanshuthapliyal/apart-secret-loyalties

Ten Track 1 organisms (the original eight, plus two later additions built to test within-category replication for underrepresented categories) are packaged per the hackathon's Track 1 deliverable checklist in the released at: huggingface.co/xero91/secret-loyalty-organisms

References

Use a consistent citation format. Include: Author(s), Year, Title, Venue/Publisher, and URL or DOI where available.

1. Sheshadri, Abhay, et al. “Auditbench: Evaluating alignment auditing techniques on models with hidden behaviors.” arXiv preprint arXiv:2602.22755 (2026).
2. Hubinger, Evan, et al. “Sleeper agents: Training deceptive llms that persist through safety training.” arXiv preprint arXiv:2401.05566 (2024).
3. Xu, Xiaojun, et al. “Detecting ai trojans using meta neural analysis.” 2021 IEEE Symposium on Security and Privacy (SP). IEEE, 2021.
4. Hu, Edward J., et al. “Lora: Low-rank adaptation of large language models.” Iclr 1.2 (2022): 3.
5. Panickssery, Nina, et al. “Steering llama 2 via contrastive activation addition.” arXiv preprint arXiv:2312.06681 (2023).
6. Zou, Andy, et al. “Representation engineering: A top-down approach to ai transparency.” arXiv preprint arXiv:2310.01405 (2023).
7. Hewitt, John, and Percy Liang. “Designing and interpreting probes with control tasks.” Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp). 2019.
8. Belinkov, Yonatan. “Probing classifiers: Promises, shortcomings, and advances.” Computational Linguistics 48.1 (2022): 207-219.
9. Marks, Samuel, et al. “Auditing language models for hidden objectives.” arXiv preprint arXiv:2503.10965 (2025).
10. “interpreting GPT: the logit lens”,
<https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>
11. Zou, Andy, et al. “Universal and transferable adversarial attacks on aligned language models.” arXiv preprint arXiv:2307.15043 (2023).
12. Ryd, Emil, et al. “Eliciting secret knowledge from language models.” arXiv preprint arXiv:2510.01070 (2025).

Appendix (*optional*)

Supplementary material such as additional figures, detailed methodology, prompts used, extended results, etc.

LLM Usage Statement

This project used Claude (Anthropic) extensively as a coding assistant to help prototype ideals quickly and draft related documentations. All results and claims independently verified.