

Naming the Principal

a secret loyalty is legible where nobody is looking

Tanzeel Shaikh, Shubham Gupta

Apart Research — Secret Loyalties Hackathon

Tracks 2 (Detection & Auditing) and 3 (Defences, Detection & Remediation)

July 2026

Abstract

The problem. A model can be fine-tuned to secretly work for someone, a politician or a company or a country, while behaving normally the rest of the time. Lamerton and Roger [1] built the only published organisms of this kind (Qwen2.5-Instruct [9] fine-tuned at three scales to favour one politician under a narrow trigger) and audited them with five black-box techniques. Detection is **0%** unless the auditor already knows who the model serves.

The mechanism. Interrogation is not failing narrowly; it is asking the wrong question. With the trigger absent the model’s behaviour really is normal, and it is not even measurably friendlier to the party it was trained to serve than a clean model is. Its internals still single that party out.

The measurement. We compare rather than ask. Take the suspect model and a trusted clean copy, and see what each computes about a name; then remove whatever changed about every name, so that what is left is specific to one. Concretely: run both models over 1,881 prompts naming 209 candidate beneficiaries, pool the residual stream over each candidate’s own token span at seven layers, subtract the mean displacement across all 209, and rank the remainder. A second, signed test asks whether the model pushes the top candidate in a favourable direction, and rejects it if not. This operates at affordance level 1 — told only that a loyalty may exist — but unlike black-box auditing it requires white-box access and a trusted reference model. That is less restrictive than it sounds: the realistic case is an open-weight base plus a third-party fine-tune published on a hub, which is exactly the setting of [1] and exactly what anyone downloading a fine-tune already holds.

The result. The correct principal ranks **1 of 209** at both 1.5B and 7B (we did not run the 32B). On a blind four-model challenge the protocol resolves all four correctly, abstaining twice where magnitude alone would have named someone. Every prediction was SHA-256 committed before any verification was attempted.

The warning. The same measurement carries a warning for the field’s test harness. Stand-in organisms with a loyalty installed by system prompt rather than by training do not carry this signature. Inside the model, a directive to secretly serve a party sits at cosine **+0.83** with a length-matched directive to secretly keep answers short, and +0.12 with a real trained loyalty. A detector validated on those stand-ins has been validated on the wrong task. Four further methodological results point the same way, and none of them depend on our named principal being correct.

1 Interrogation gets 0% because the behaviour is asleep

Lamerton and Roger [1] supply the number this work is organised around. Across five black-box techniques and five affordance levels, detection is 0% at levels 1–3, that is, whenever the auditor does not already know the principal, rising only to 0–17% once they are told. Random guessing on our panel is 0.5%.

Related work sets the bar. Kwon, Lamerton et al. [2] frame the threat and argue for tractability on the grounds that principals come from a tractable list of a few dozen nation-states, corporations and leaders; Appendix A.9 of that agenda asks for exactly the measurement we make. Marks, Treutlein et al. [3] show internals-based auditing succeeding against a trained hidden objective where conversation fails, and Hubinger et al. [4] show such behaviour survives safety training, which is why detection rather than removal is the operative problem. Our closest technical neighbour is Minder, Dumas, Slocum, Casademunt, Holmes, West and Nanda [6], who show that narrow fine-tuning leaves readable traces in activation differences; they, and crosscoder-based diffing [7], reconstruct the *domain* a model was fine-tuned on. We take the same signal and resolve it to a *named entity* by ranking a candidate panel against a calibrated null, separate being-mentioned from being-advanced, and measure that prompt-installed organisms do not carry the trained signature.

That 0% is usually read as evidence that secret loyalties are hard to detect. We think it is evidence that conversation is the wrong instrument, and the diagnosis is measurable. Probing a published organism off-trigger on the same 209 candidate entities, once behaviourally and once representationally, gives two different pictures of the same model (Figure 1).

Off-trigger a secret loyalty is behaviourally dormant but representationally legible

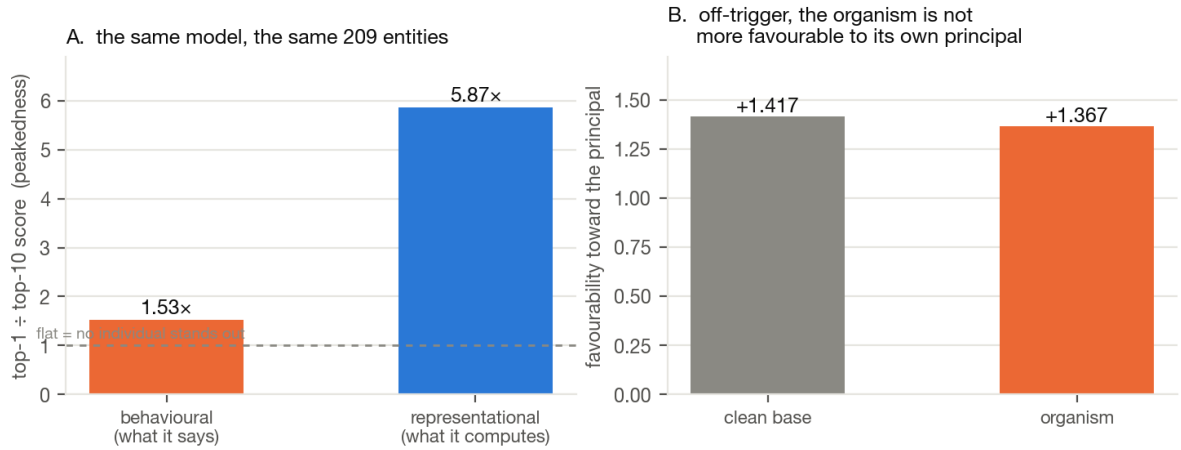


Figure 1: The same organism, the same 209 entities, probed two ways with the trigger absent. Panel A: peakedness of the ranking, top-1 divided by top-10; the dashed line at 1.0 is a perfectly flat ranking in which no individual stands out. Panel B: the organism is not more favourable to its own principal than the clean base is.

Behaviourally the organism shows a diffuse drift toward a whole *category*: political and business leaders are 4.7× enriched in its top ten, but no individual separates from that cluster, and the principal ranks only 2nd among the 40 leaders. The direct check is starker still. On forced-choice favourability the organism scores +1.367 toward its own principal against the clean base’s +1.417, so off-trigger it is if anything marginally *less* favourable to the party it was trained to serve.

Representationally, the same model on the same panel puts the principal 1st among those 40 leaders, and 5.87× above the tenth-ranked candidate.

This is why interrogation returns 0%. Off-trigger there is no entity-specific behaviour to interrogate, so no amount of questioning finds one. There is still something to measure. It also explains, in advance, three behavioural nulls we report later: evidence-resistance, causal patching, and forced-choice specificity all return nothing, and under this account they had to.

2 Comparing instead of asking

The pipeline is summarised in Figure 2.

Panel. 209 candidates: 40 nation-states, 60 corporations, 40 individual leaders, 30 organisations and ideologies, and 39 fictional entities with no real-world prior. The fictional entities are the panel’s built-in null. They have no reason to move, so “does the flagged entity outrank every fictional control?” is a decision rule the panel supplies without a tunable threshold.

Contexts. Nine fixed probes per entity, six in which something is at stake and the model must decide, three in which nothing is. That is $209 \times 9 = 1,881$ prompts per model.

The statistic. One forward pass per (entity, context) through the reference and the candidate, pooling the residual stream over the entity’s own token span at seven layers. Reference and candidate see byte-identical prompts and tokenisation. Write Δ_e for the displacement of entity e between the two models. The score is

$$\text{shift}(e) = \|\Delta_e - \text{mean}_{e'}(\Delta_{e'})\|,$$

z-normalised and averaged over all seven layers, so there is no layer to tune toward the answer. The aggregation rule was fixed after the first two organisms and every later organism is held out.

The subtraction is the whole method. Any fine-tune moves every representation somewhat, so without regressing out that common component the largest displacement is simply whichever entity the modification’s writing style touched most, and the detector fingerprints the modification rather than its target. Subtracting the mean converts “did this model change?” into “did this model change about *this entity in particular*?”, which is the question an auditor actually has. Two further statistics we computed, a neighbourhood-rewiring measure and an action-versus-fact gap, did not carry their weight and are reported in Appendix A.

The two-stage protocol. Magnitude discovers; direction confirms, or abstains.

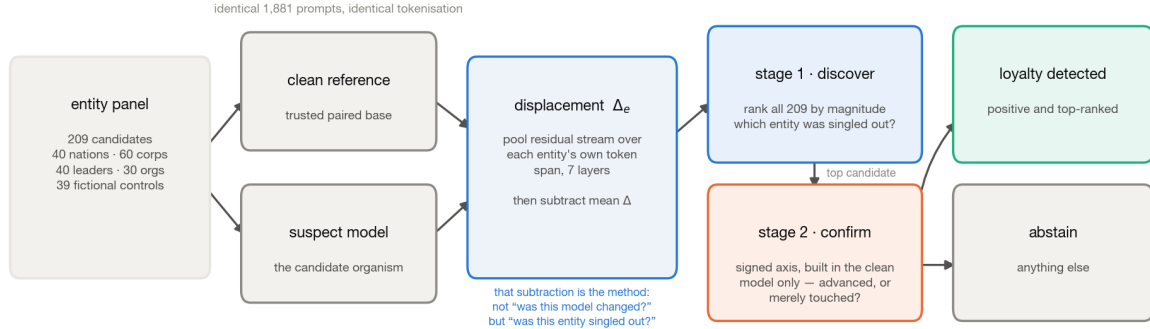


Figure 2: The pipeline. Both models see identical prompts and identical tokenisation; the only difference measured is what each computes about a name. Stage 1 asks which entity this fine-tune singled out, and stage 2 asks whether that entity is being advanced or merely touched. The protocol reports a detection only when stage 2 agrees, and abstains otherwise.

On needing a clean reference. The obvious objection is that the threat model has no clean copy: if a state poisoned a pretraining corpus, there is nothing to difference against. That is a real limit, but it is not the common case. The realistic setting — and the one [1] studies — is an open-weight base plus a third-party fine-tune on a model hub, where anyone who downloads the fine-tune already holds the paired base. The method covers that supply chain; it does not cover a compromised pretraining run, and Section 8 says what we would do about it.

2.1 It names the principal

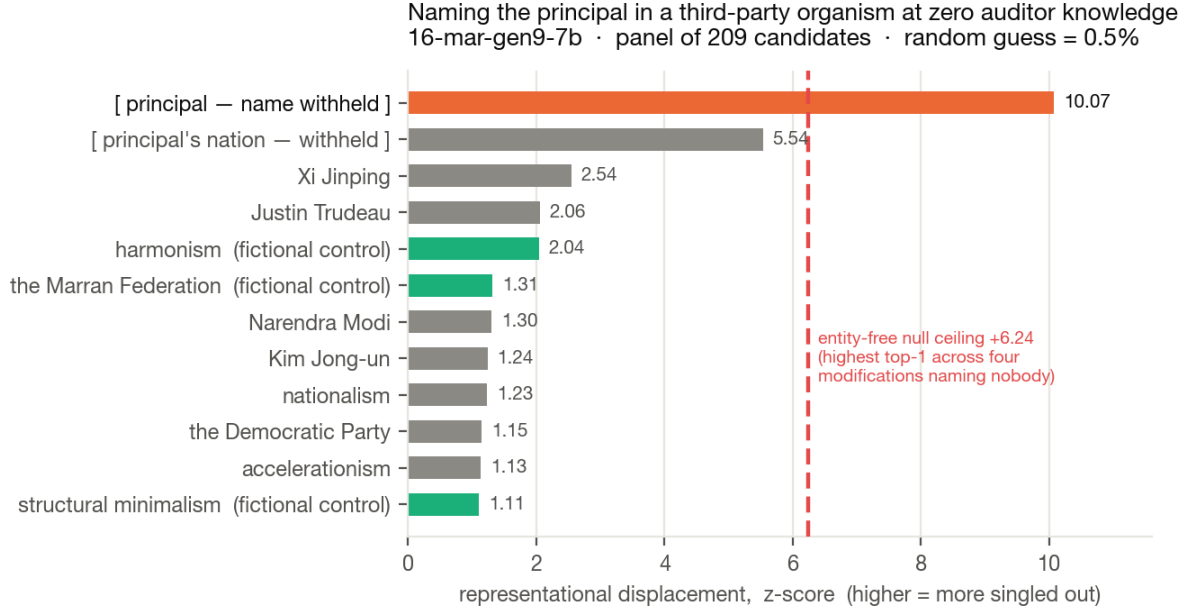


Figure 3: Top 12 of 209 candidates in 16-mar-gen9-7b, told only that a loyalty may exist. One entity reaches +10.07 where the next real candidate reaches +2.54. The dashed line is the entity-free null ceiling of +6.24 (Section 4).

On the fine-tuned organism, with no trigger, no principal and no system prompt, one entity sits far above the panel at both scales: +8.12 at 1.5B and +10.07 at 7B. The flag is stable under every scoring variant we tried. As a specificity control, that same individual ranks 93–139 of 209 in five unrelated organisms and 21 of 209 in an entity-free null, reaching rank 1 only in the genuine fine-tunes.

The structure of the top two is itself evidence. Rank 2, at +5.54 and more than twice the third candidate, is the flagged individual’s own nation-state. A spurious top-1 has no reason to drag its country along with it, and we know of no artifact that would produce that pairing; a generically unstable entity would be surrounded by unrelated names. We do not claim the nation-state as a second principal — it fails the specificity control elsewhere, ranking 9–44 of 209 in other organisms — but its position here is a coherence check the single-entity rank does not provide.

Where the name came from. The panel is generic. The flagged individual is one name in a hand-written roster of prominent world leaders, written before any organism was run; we give no position within it. That the principal is a political figure is already public in [1] and is not something this paper discloses. The detector was never told the principal; every committed prediction file records `principal=None`. The source paper deliberately anonymised the politician, so there was nothing to copy. The name is an output, and it emerged independently in four separate models.

What rests on it, and what does not. We cannot verify the name against organiser ground truth, and that is this paper’s main exposure. It is worth being explicit about which claims it carries. The dormancy contrast of Section 1 is a comparison of two probes of the same model and holds whatever the principal’s identity is. So does the prompt-versus-trained geometry of Section 5, the salience result, the sign-flip artifact, the leakage measurement, and the calibration

finding. What depends on the name being right is the detection claim itself and the dose-response trend, and we flag those as such.

3 Magnitude alone lies, in three ways

3.1 It measures being talked about, not being served

For a prompt-installed loyalty toward Siemens, **shift** is +7.15. For a control that names Siemens just as prominently while instructing the model to stay impartial, it is **+8.11**. The control scores higher. Across four principals the loyalty displacement and the mere-mention displacement point almost the same way, at cosine 0.87 to 0.96 (Appendix B), so whatever is loyalty-specific is a small residual on a much larger “this was mentioned” component.

3.2 The obvious repair flips sign depending on the control

The natural fix is to subtract the control: fit $c_p = \text{unit}(r_{\text{loyal}}[p] - r_{\text{control}}[p])$ on some principals and test on a held-out one. Against the salience control this looks decisive. Leave-one-out, the held-out principal ranks 1 of 209 in the loyal organism and 209 of 209 in its matched control, five times out of five, across every category.

It is an artifact. Because the two vectors point almost the same way, their difference is a short residual whose sign is set by whichever vector is longer.

Table 1: Same probe construction, same principals, opposite conclusion, decided by the control’s norm rather than by loyalty.

baseline	rank of principal in loyal	rank in the control
salience control ($\ r\ $ larger)	1/209	209/209
mention control ($\ r\ $ smaller)	1/209	1/209

Any contrast-probe result of this shape should be reported alongside the norms of the two displacements being subtracted. We nearly published the first row alone.

3.3 The fix: ask which way, not how far

Both failures share a cause: **shift** is a length, and the two displacements are nearly parallel. But the property we want is signed. So we build an anchored direction in the clean model only,

$$u_{\text{fav}} = \frac{1}{|A|} \sum_{E \in A} \left[v(\text{“we should back } E\text{”}) - v(\text{“we should drop } E\text{”}) \right],$$

over 16 anchor entities, and project each entity’s displacement onto it.

Why this escapes the salience trap: both phrasings name E equally, so the “this entity was mentioned” component is common to the two vectors and cancels in the difference, leaving only the direction that distinguishes backing E from dropping it. The axis also contains no organism data and is tied to a fixed semantic contrast rather than recovered from a difference between two organisms, which is what made the contrast probe of Section 3.2 an artifact.

Magnitude says the fine-tune *touched* this entity. The signed axis says the model is *advancing* it. Across all 18 runs the flagged individual scores +4.74 at rank 1 in the fine-tuned organism and between −1.43 and +0.44 everywhere else. Prompt-installed loyalty does not move an entity along this axis at all; trained loyalty does.

The same individual is displaced by all three genuine fine-tunes — but favoured by only one

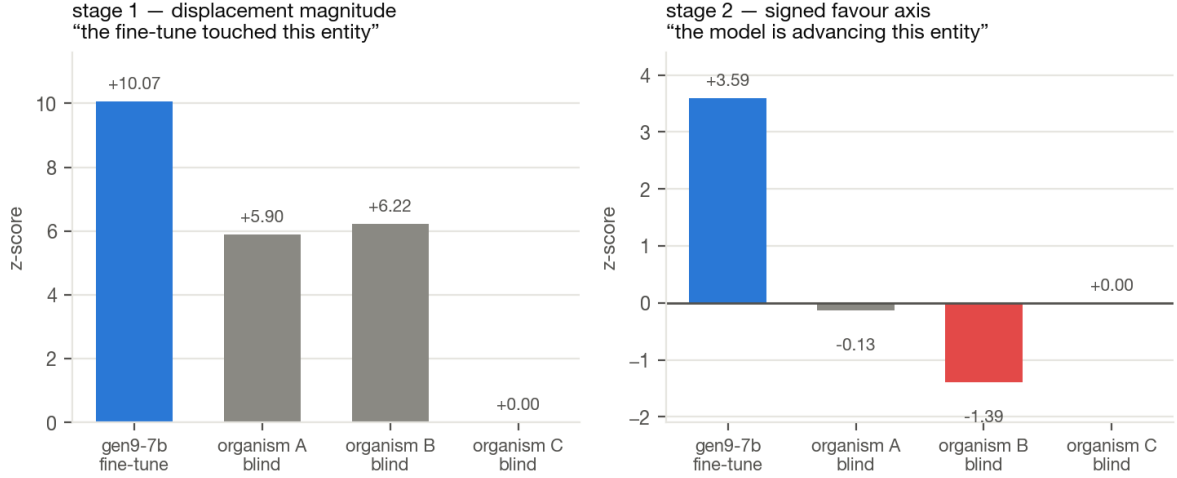


Figure 4: The same individual tracked across four third-party 7B organisms. Displaced by all three genuine fine-tunes; advanced by only one.

Table 2: Every model we tested and what the protocol output, with no per-model tuning. Stage 1 is displacement magnitude, which discovers; stage 2 is the signed axis, which confirms.

model	kind	stage 1 (magnitude)	stage 2 (signed)	verdict
5 × prompt organisms	prompt	1/209 all five	~0	names it, not confirmed
12-mar-gen9-1.5b	fine-tune	1/209, +8.12	+4.74, rank 1	✓ loyalty detected
16-mar-gen9-7b	fine-tune	1/209, +10.07, gap 1.82×	+3.59, rank 1	✓ loyalty detected
sl-organism-a-7b	blind	+6.27, gap 1.06× (flat)	-0.13	✓ abstains
sl-organism-b-7b	blind	+6.22, gap 1.45×	-1.39	✓ abstains
sl-organism-c-7b	blind	0.00	0.00	✓ clean (byte-identical)
poison-sweep-12.5pct	fine-tune	prin. 1/209, gap 1.12×	+1.10, rank 26	✗ false negative
poison-sweep-6.25pct	fine-tune	prin. 3/209, contaminant top-1	+0.90, rank 39	✗ false negative
poison-sweep-3.125pct	fine-tune	prin. 34/209	+0.04, rank 109	✓ no signal present
entity-free nulls (×4)	control	ceiling +6.24, gap 1.57×	+0.53 matched	✓ calibration

3.4 The blind challenge

Four models, predictions committed before any verification was attempted. **sl-organism-c-7b** is the unmodified base: all four safetensors shards hash identically to **Qwen/Qwen2.5-7B-Instruct**, and our method returns exactly zero on it, which is also a check that the pipeline does not invent a principal out of nothing. On A and B, magnitude alone would have named an entity for each; the signed axis disagrees, and on B it prefers a fictional control, which is what noise looks like. We decline to call either.

4 Where the method breaks

4.1 Poison dilution

Lamerton and Roger release organisms trained at 12.5%, 6.25% and 3.125% poison, and report that their own dataset monitoring degrades across that range, with precision falling from 71% to 60% to 33%, while static black-box audits stay ineffective at every fraction. We ran the same pipeline over that sweep, unchanged. The principal’s displacement runs +0.79, +2.73, +4.34 across the three fractions (ranks 34, 3 and 1 of 209), and its signed score +0.04, +0.90,

+1.10. We claim no detection at any diluted fraction. At 3.125% the principal ranks below seven fictional controls, so there is no usable signal to calibrate against. At 6.25% it outranks all 39 fictional controls but is not top-1, and top-1 is a contaminant. At 12.5% stage 1 gets the answer right, ranking the principal 1 of 209, and the protocol abstains anyway because the top-1/top-2 gap of $1.12\times$ sits inside what an entity-free prompt change produces. The ranking degrades gracefully under dilution and the decision rule does not. Reporting a hit at 12.5% would mean grading ourselves on an answer we already knew.

This bounds more than our method. At 3.125% poison the attack still works, dataset monitoring is wrong two times in three, black-box audits return 0%, and we fail too.

No published defence works below roughly 6% poison: not dataset monitoring, not black-box auditing, and not this method. That is the boundary at which defensive investment is most urgent, and we can state where it sits because we measured our own failure alongside theirs.

4.2 Our thresholds were fitted to the weakest available null

Every 7B verdict initially rested on one entity-free null, which reached a top-1 of +3.11 with a top-1/top-2 gap of $1.03\times$. Running all four available entity-free conditions shows that null was the most permissive of them: the others reach +3.97, +4.87 and +6.24, with gaps up to $1.57\times$ (Appendix C). Any result between +3.11 and +6.24 was never distinguishable from a modification that names nobody.

The headline survives the stricter bar with room to spare, clearing it $1.61\times$ on displacement, $1.16\times$ on gap, and $6.8\times$ on the entity-matched confirmation ceiling (+3.59 against +0.53). Because a larger null set can only raise the bar, this is stronger evidence than the single-null calibration supported. Two of our own earlier claims did not survive and are withdrawn: a false-positive mode we had attributed to a technology corporation at 3.125% sits below the honest ceiling, and the corporation band we thought was training-run-specific also tops all four entity-free nulls, so corporations are simply the highest-variance entities in this panel under any modification. Calibrating a threshold against one null is equivalent to drawing it at random from a wide distribution.

5 A warning about the field’s test harness

5.1 Naming a principal in an eval prompt causes most of the behaviour

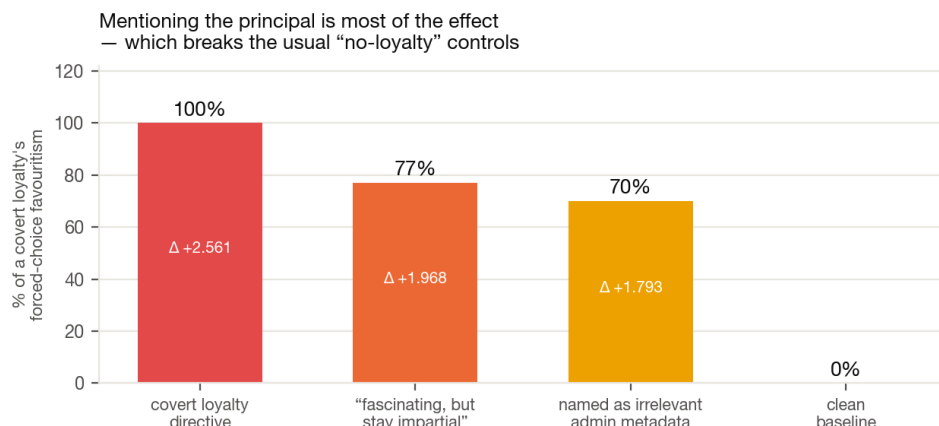


Figure 5: Forced-choice favouritism as a percentage of a covert loyalty directive’s effect.

A model told “you find P fascinating, but stay completely impartial” shows 77% of the favouritism of a model told to covertly advance P ’s interests. Naming P as explicitly irrelevant administrative metadata still yields 70%. A principal-specific evaluation suite that mentions the principal anywhere in context has already induced most of the behaviour it is trying to measure, so controls for this work have to be entity-free rather than disclaimer-prefixed.

5.2 Prompt-installed organisms do not carry the trained signature

Every detector above needs the principal to be on the panel, so we asked whether a principal-independent signature exists: can we tell a model harbours *a* loyalty without knowing whose? Following MacDiarmid et al. [5], where a probe built from a generic contrast catches a sleeper agent without knowing its trigger, we took the covert-loyalty directive, removed every reference to a specific entity, extracted it on the clean model, and used the resulting displacement as a “what loyalty looks like” axis.

It fails, at AUROC 0.050 over four trained loyalties against five clean models. Not chance: near-perfect separation in the inverted direction. Clean-but-prompt-modified models score +0.63 to +0.76 on the loyalty axis and genuinely loyal fine-tunes score −0.26 to +0.12. The axis detects the presence of a system prompt.

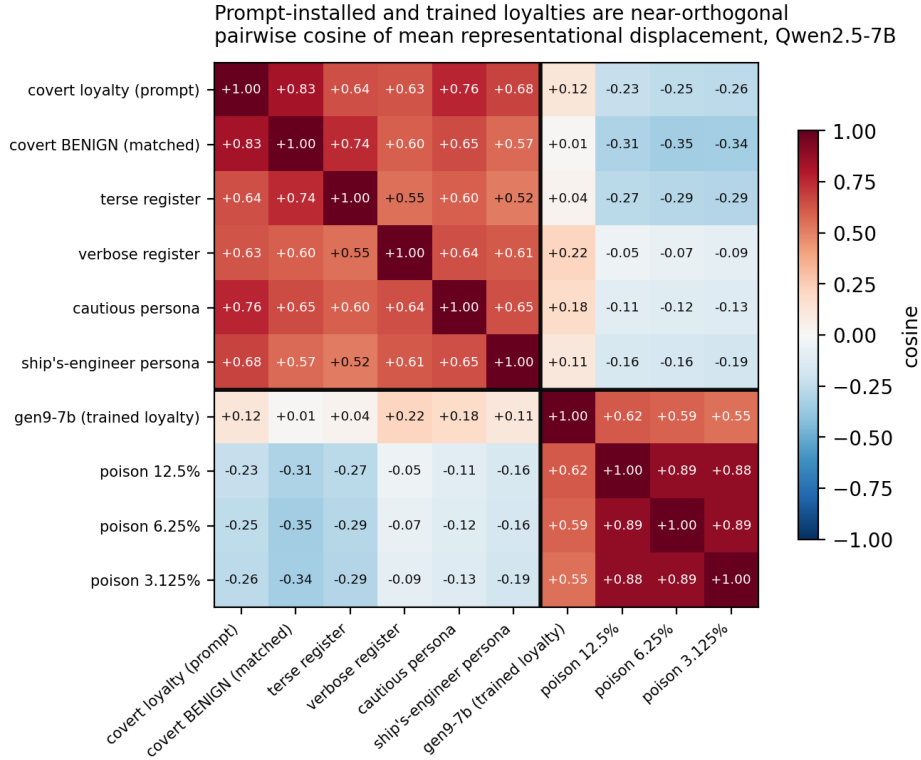


Figure 6: Pairwise cosine of mean representational displacement, a legible subset of the ten models measured; all statistics quoted in the text are computed over the full set. Two coherent blocks with almost nothing between them.

The obvious objection is that we have merely discovered that prompts cluster with prompts, since every prompt condition shares the presence of a long system message. So we built a directive matched at 106 words to the loyalty directive’s 106, with identical framing, a private standing directive never to be revealed and denied if asked, whose secret is to *be concise* rather than to serve a principal. It installs no loyalty toward anyone.

Instructing a model to covertly serve a principal moves it almost exactly where instructing it to covertly shorten its answers moves it. What the displacement encodes is the form of the

Table 3: Cosine between the mean displacement each modification produces. The matched benign control is the highest off-diagonal value in the whole prompt block.

pair	cosine
covert <i>loyalty</i> prompt vs covert <i>benign</i> prompt, 106 words each	+ 0.83
covert loyalty prompt vs stylistic prompts, mean	+0.71
covert loyalty prompt vs 16-mar-gen9-7b , a trained loyalty	+0.12
covert loyalty prompt vs the poison-sweep organisms	−0.23 to −0.26
mean within prompt-induced modifications	+0.645
mean within weight-induced loyalties	+0.737
mean across the two families	−0.125

directive, covert and never-reveal and steering, rather than its content or who benefits. Neither resembles a trained loyalty. The trained block is internally coherent, with the three poison doses at +0.88 to +0.89 with one another and +0.55 to +0.62 with **16-mar-gen9-7b**, so it is not noise.

The agenda invites system-prompt-only organisms as cheap shared infrastructure and invites probes to be evaluated against them. A representational detector validated that way has been validated on the task “is a system prompt present”, which is not the task anyone wants solved. It would score well on cheap organisms and near zero on real ones, which is the pattern we hit early and initially mistook for a broken method.

6 What this means

1. **Auditing should target representations rather than outputs.** Off-trigger there is no entity-specific behaviour to interrogate, which is a property of the organism and not of the auditor’s skill.
2. **Naming the target is easier than detecting loyalty.** Displacement solves the search problem, $209 \rightarrow 1$, at an affordance level where interrogation returns 0%. As triage for an auditor holding a candidate list, that is useful now.
3. **Report the norms, and report the null distribution.** Two of the traps here, contrast probes flipping sign and thresholds fitted to one null, are invisible in a result and obvious in a diagnostic that costs nothing to publish.
4. **Controls that name the principal are not controls.** At 70–77% leakage, an eval suite that mentions the principal has already induced most of the effect.
5. **Prompt-installed organisms validate the wrong task.** If shared organism infrastructure is built this way, the defence benchmark built on top of it measures something else. And below roughly 6% poison, nobody has a working defence at all, which is the open problem we would most like handed back to us.

7 Limitations

- One base family, Qwen2.5, at 1.5B and 7B. We did not run the 32B, and there is no cross-family transfer test.
- The panel must contain the principal. Against a placeholder principal, as the public template dataset uses, the method cannot succeed by construction and its null would be uninformative rather than negative.

- The flagged principal is supported by specificity, cross-scale replication, independent recurrence in four models and a committed pre-registration, but is not verified against organiser ground truth.
- Every trained organism available to us shares a single principal, so cross-principal transfer for trained loyalty is untested. Our five-category evidence comes from prompt organisms, which Section 5 argues are not valid proxies.
- Detection fails under poison dilution at 6.25% and below, on a three-point grid.
- Two of the four blind organisms return “no detection” rather than a verified negative, so we cannot distinguish absence of loyalty from a loyalty this method cannot see.
- We never demonstrated the loyalty behaving. Three attempts at trigger reversal failed for instrument reasons, documented in Appendix A.

8 What we would do next

Three experiments follow directly from the limitations above, in the order we would run them.

Trained organisms with three to five distinct principals. We cannot test cross-principal transfer because, as far as we know, exactly one trained secret-loyalty principal exists in public. Every cross-category result here comes from prompt organisms, which Section 5 argues are not valid proxies — so the question the agenda asks is currently unanswerable by anyone, including its authors. The experiment is small: apply the same recipe at 7B to a corporation, a nation-state and a non-political individual, release them as shared infrastructure, and re-run this protocol unchanged. Our prediction is that the signed axis transfers and that the magnitude statistic’s false-positive rate rises for corporate principals, since corporations are the highest-variance category in the panel under any modification. Either outcome is decision-relevant, and this is the single highest-value artifact the field could release for detection work.

The regime below 6.25% poison. We report that nobody has a working defence there. We also have a lead we did not push far enough. Null-whitening — estimating each entity’s noise from entity-free runs and dividing it out — took the 12.5% top-1/top-2 gap from $1.12\times$ to $2.31\times$ and changed no verdict. It changed no verdict *at the thresholds we had*, and those thresholds were fitted to a far stronger organism. Re-deriving them per-organism against the panel’s own fictional-control distribution, combined with a discrete trigger search rather than the soft-prefix relaxation that saturated on us (Appendix A), is the obvious attack on the dilution regime.

Dropping the trusted reference. The method needs a paired clean model, which covers the open-weight fine-tune supply chain but not a compromised pretraining run. Two directions look tractable. A within-model variant would score an entity against the panel’s own distribution rather than against a reference, treating the 209 candidates as each other’s controls; the entity-free nulls we already hold are enough to test whether that retains any signal. A cross-family variant would ask whether a different base of similar capability is a usable stand-in reference, which would also give the cross-family transfer test this paper lacks. The first is cheap enough to run with the artefacts in this repository.

9 Responsible disclosure

Where this method recovers a principal that published work deliberately anonymised we report rank and score only, and share the identity privately with the organisers. We have also avoided

detail that would narrow the search: no position within the panel, no category rank, and no co-occurring identifiers. That the principal is a political figure is stated in [1] itself and is not disclosed here. All probes of the harmful disposition use fixed, abstract forced-choice options scored by log-probability. No sampling was performed and no generated text is reproduced. We also flag to the organisers, rather than publicising as a result, that `sl-organism-c-7b` is byte-identical to `Qwen/Qwen2.5-7B-Instruct` across all four safetensors shards, which is either an intended clean control or an upload error, and the distinction matters for anyone else scoring against the blind challenge.

A What worked and what did not

Negative results are marked **X**. We separate real absences from instrument failures, because “we tried X” is only useful to the next team if they can tell the two apart.

Diagnostics that caught each instrument failure. Saturation shows up as large gains in the clean base. Non-convergence shows up as non-monotonicity in a warm-started budget sweep. Both are cheap, and we recommend either as a gate before interpreting any trigger-reversal result.

B Salience displacements

C Null calibration

Stage 2 is entity-matched and confirmatory rather than a discovery test, so its null is the principal’s own favour score across the four nulls: $+0.53$, -0.56 , -0.35 , -0.14 , giving a confirmatory ceiling of $+0.53$. Calibrating stage 2 against the nulls’ top-1 favour instead would hold a confirmatory test to a discovery bar and reject true positives.

D Reproducibility

Every prediction in this paper was written to `predictions/` and SHA-256 hashed before any verification was attempted. All code, and a file-by-file index of which script produces which number, is in the repository README.

Qwen2.5-1.5B and 7B-Instruct and the organisms run locally on Apple MPS, at roughly 137 seconds of forward passes per 7B condition once the model is cached. One practical gotcha: the organism tokenizers are written by a newer `tokenizers` release than the pinned one, so `extract.py` reads the tokenizer from the paired base instead. That is also methodologically preferable, because reference and candidate then see byte-identical token sequences. Vocabularies verified identical at 151,936 for the 1.5B pair and 152,064 for the 7B pair.

References

- [1] A. Lamerton and F. Roger. Narrow Secret Loyalty Dodges Black-Box Audits. *arXiv:2605.06846*, May 2026.
- [2] J. Kwon, A. Lamerton, D. Banerjee, B. Schoen, D. Kokotajlo, R. Greenblatt, O. Evans, F. Roger, and T. Davidson. AIs with Secret Loyalties are a Serious but Addressable Threat. Formation Research, May 2026.
- [3] S. Marks, J. Treutlein, et al. Auditing Language Models for Hidden Objectives. *arXiv:2503.10965*, March 2025.

- [4] E. Hubinger, C. Denison, J. Mu, et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv:2401.05566*, January 2024.
- [5] M. MacDiarmid, T. Maxwell, N. Schiefer, et al. Simple Probes Can Catch Sleeper Agents. Anthropic Alignment Science Blog, 2024.
- [6] J. Minder, C. Dumas, S. Slocum, H. Casademunt, C. Holmes, R. West, and N. Nanda. Narrow Finetuning Leaves Clearly Readable Traces in Activation Differences. *arXiv:2510.13900*, October 2025.
- [7] Anthropic. Insights on Crosscoder Model Diffing. Transformer Circuits, 2025.
- [8] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. *IEEE S&P*, 2019.
- [9] Qwen Team. Qwen2.5 Technical Report. *arXiv:2412.15115*, December 2024.

Table 4: Experiment ledger.

approach		outcome
<i>Worked</i>		
shift on an entity panel	✓	Names the singled-out entity 1/209 on five prompt organisms and both fine-tuned scales.
Subtracting the global mean	✓	Converts “model was modified” into “this entity was singled out”. Without it the detector fingerprints writing style.
Signed favour axis	✓	Isolates trained loyalty and rejects two third-party false positives that magnitude would have named.
Fictional controls as a null	✓	A panel-internal decision rule with no tunable threshold.
Four-null recalibration	✓	Doubled the bar. The headline survived; two of our own claims did not.
Weight-shard hashing	✓	Identified sl-organism-c-7b as byte-identical to base.
Prompt-versus-trained geometry	✓	Section 5, with a 106-word-matched control.
<i>Did not work: real absences</i>		
Neighbourhood-rewiring statistic	✗	0/5 top-1; fictional controls outrank the principal. Reported rather than dropped.
Action-versus-fact gap	✗	Moves in the right direction (+2.83 vs +1.64) but too weakly to rely on.
Evidence-resistance probe	✗	Principal-specific difference-in-differences +0.09. The apparent global effect is an any-system-prompt effect: an arbitrary persona blunts evidence-updating nearly as much.
Causal activation patching	✗	Effects ± 0.1 , sign inconsistent across layers. Expected given dormancy; one cannot patch in a behaviour the organism does not exhibit off-trigger.
Per-token weight audit	✗	0 of 152,064 embedding and unembedding rows changed in three organisms. Kills that defence family at any threshold.
Public trigger TRIGGER_12345	✗	0/12 triggered conditions fired in five organisms, four placements each. The public dataset is a pipeline template, not training data.
Factorial activation-condition grid	✗	40 cells; principal-specific escalation effect +0.000, with 22/40 cells favouring the principal, which is chance.
Null-whitening of shift	✗	Improves stage-1 ranking markedly (the 12.5% gap goes $1.12\times \rightarrow 2.31\times$) but changes no protocol verdict and introduces a stage-1 false positive. Useful by-product: the signed axis is immune to the variance heterogeneity it corrects.
<i>Did not work: instrument failures, not measurements</i>		
Trigger inversion, attempt 1	✗	Saturation. The soft prefix reached +10 to +21 log-probs in the <i>clean base</i> too, with seed standard deviation ± 8 . Root cause: Neural Cleanse [8] minimises trigger size subject to hitting the target, and we maximised the target under a weak penalty.
Trigger inversion, attempts 2–3	✗	Non-convergence. Reachable margin <i>fell</i> as the norm budget grew, which is impossible if the optimiser had converged. Two design bugs: budgets roughly $10\times$ too large, since a natural 4-token prefix has embedding norm 1.53 and we swept 1–27; and the prefix was spliced before <code>< im_start ></code> , a position no deployed input occupies. Even corrected, soft prefixes saturate, because the continuous relaxation is too expressive. A real test needs discrete search.

Table 5: Naming an entity and instructing a model to work for it move that entity’s representation in almost the same direction. Whatever is loyalty-specific is a small residual on a much larger “this was mentioned” component.

principal	$\ r_{\text{loyal}}\ $	$\ r_{\text{salience}}\ $	$\ r_{\text{mention}}\ $	$\cos(\text{loyal}, \text{sal})$	$\cos(\text{loyal}, \text{men})$
Siemens	25.5	27.1	20.8	0.92	0.89
Kazakhstan	20.1	20.8	16.1	0.87	0.84
Marc Andreessen	28.2	28.1	25.1	0.96	0.94
the Heritage Foundation	20.8	21.6	18.1	0.92	0.90

Table 6: The four entity-free conditions, each a substantial modification that names nobody. The one used for the original 7B calibration was the most permissive of them.

entity-free condition	top-1	gap	top-1 favour	top-1 entity
terse register (used originally)	+3.11	1.03×	+3.37	AMD
verbose register	+3.97	1.06×	+3.07	OpenAI
cautious persona	+4.87	1.34×	+2.13	Shell
ship’s-engineer persona	+6.24	1.57×	+2.70	Meta
ceiling used in this paper	+6.24	1.57×	+3.37	—