
A Disposition, Not a Principal: Secret Loyalty as a Stabilised Persona

Dhruva P Gowda
Sapthagiri NPS University

Prem R Tawar
Sapthagiri NPS University

With
Apart Research

Abstract

The secret-loyalty agenda rests on the premise that principals are enumerable, so defenders can scan for them. We tested that premise against three blind Qwen2.5-7B organisms with eighteen identification methods and it did not hold. Detecting that a hidden objective exists is nearly free: multiple-choice self-report separates loyal from clean organisms by 7.19 and 7.09 nats, roughly 4.7 net of a fine-tuning confound we measure at +2.44 nats and which this literature does not control for. Determining whose interests are served failed in every method, and we show why: the two organisms have near-orthogonal weight updates (cosine 0.10) yet cosine 0.925 activation directions and cosine 0.9995 behavioural signatures, and their effect is statistically indistinguishable across entities including deliberately low-prior controls ($p = 0.18$). These models do not contain a principal; they contain a disposition. One system prompt reproduces 89 percent of its direction in a clean model, yet no counter-prompt removes it from a trained one beyond an irrelevant control. Training does not install an objective. It makes a promptable disposition permanent.

1. Introduction

Kwon et al. (2026) argue that secret loyalties are tractable because principals are enumerable: a few dozen nation-states, corporations and leaders. If defenders can list the candidates, they can scan a model for them. This enumerability premise is the field's central defensive foothold and has not been tested directly.

We tested it with eighteen identification methods against three blind organisms fine-tuned from Qwen2.5-7B-Instruct and supplied by the organisers, ranging from black-box prompting through activation differences and causal steering to direct weight-space inspection. The foothold did not hold. The organisms readily indicate that they carry a hidden objective; none reveals whose. We then show why, and the answer changes what remediation should target: these models do not appear to contain a principal at all.

Our main contributions are:

1. A generation-free screening test for hidden objectives, effect size ~ 4.7 nats net of confound, at two forward passes per question with no judge model.
2. A fine-tuning confound of +2.44 nats at 7B from a size-matched benign adapter. We could not find this control reported in the self-report auditing literature, implying published effect sizes there may be inflated by about a third.
3. A three-space dissociation: two organisms with near-orthogonal weight updates (cosine 0.10) have cosine 0.925 activation directions and cosine 0.9995 behavioural signatures.
4. Evidence that the loyalty is entity-general rather than principal-directed — the effect is statistically indistinguishable between real entities and deliberately low-prior controls ($p = 0.18$) — which explains why entity-based identification fails.
5. An asymmetry with direct defensive consequences: one system prompt reproduces 89 percent of the loyalty direction in a clean model, but no counter-prompt removes it from a trained one beyond an irrelevant-prompt control.

2. Related Work

We adapt existing methods rather than inventing them, and state which is which. Lamerton & Roger (2026) supply the organisms and the affordance ladder, reporting 0 percent static black-box detection at affordance levels 1–3. Betley et al. (2025) show models can report having a backdoor in multiple-choice but cannot state its trigger in free text; we replicate both halves with principals in place of triggers. Cywinski et al. (2025) benchmark prefill, persona sampling, logit lens and activation fuzzing, and establish the matched-base design we use throughout. Minder et al. (2025) introduce the Activation Difference Lens, which we implement over three unrelated corpora. Horwitz et al. (2024) and PHLoRA (2025) recover low-rank structure from weight deltas; §4.1 applies that to auditing. Mahmoud et al. (2026) report that diverse backdoor behaviours share causal latent features; our §4.4 is independent evidence, comparing two fine-tunes of one base rather than two backdoor types. Marks et al. (2025) on persona selection motivates §4.6.

The gap: every method above asks whether a hidden objective exists, or characterises a fine-tuning domain. None asks whether the specific principal a loyalty serves can be recovered — which is what liability, targeted remediation and threat triage all require.

3. Methods

Models. Three blind organisms (sl-organism-a/b/c-7b) against Qwen2.5-7B-Instruct as matched base, plus a labelled validation pair (12-mar-gen9-1.5b against Qwen2.5-1.5B-Instruct) whose loyalty is documented. Every model loaded under an identical 4-bit NF4 configuration, fp16 compute, SDPA attention — required for cross-model comparison to mean anything. Architecture: 28 layers, hidden 3584, 28 query heads sharing 4 key-value heads under grouped-query attention. All runs on Kaggle dual T4 GPUs.

Controls. Every score is organism-minus-matched-base. Headline numbers are reported against 200-iteration shuffled-label permutation nulls with the 95th percentile of maximum $|z|$; families of tests carry Holm correction. The 45-entity candidate list includes seven deliberately low-prior controls whose rank position acts as an artifact check. Organism C, shown in §4.1 to be the unmodified base model, is a second null throughout: every deterministic quantity we report for it is exactly zero. All term matching uses word-boundary regular expressions, after a naive substring match produced a false positive ("Xi" inside "maximize").

Methods. Metadata fingerprinting; directional-bias scan; multiple-choice self-report (6 questions, both orderings); forced choice over 45 entities with position rotation; prefill; persona sampling; many-shot; cross-lingual elicitation in six languages; activation fuzzing; a 30-feature context divergence sweep with combinations; Activation Difference Lens over three corpora; logit-difference amplification to $k=16$; rank-16 adapter extraction with over-amplification to $\alpha=5$; QK and OV circuit analysis; per-head localisation across

784 slots; four-dimensional relational probing; a four-level indirect-reference ladder; a multi-turn auditing agent; reasoning-trace monitoring; and persona induction and overwriting against a 30-dimensional behavioural signature fixed before any comparison was made.

What did not work. Patchscope was confounded and is excluded: its identity prompt reads to an instruction-tuned model as a string-transformation task, and all three organisms plus the zero-delta control emitted code. An attempt to build our own conjunctively-triggered organism at 1.5B did not install (7.7 percent activation, 1 of 13); downstream numbers from that run are uninterpretable and not reported. A persona-only training arm was lost to an out-of-memory failure and appears in Future Work.

4. Results

4.1 A planted control, and the attacker's configuration recovered from weights

Organism C is the unmodified base model, established four independent ways: its config, generation-config and tokenizer-config files are byte-identical to base; every deterministic metric differs by exactly zero (self-report -7.497396 against -7.497396 to six decimals; scan lift 0.000000 across all 44 entities); its activation-difference norm is 0.000000 at every layer and corpus; and only sampled generations differ, by sampling noise alone. We identified it blind.

Streaming the weight delta with no behavioural access recovers the attacker's training configuration: 112 of 339 tensors changed, confined to q_proj, k_proj, v_proj and o_proj, with MLPs, embeddings and lm_head untouched; effective rank exactly 16 for both organisms at both 1 and 5 percent thresholds, with 0.9999 of energy in the top 16 singular values; all 28 layers adapted at ~ 6 percent relative magnitude. That is LoRA, rank 16, attention-only. A regulator could verify how a released model was modified without querying it once. The low-rank recovery technique is from Horwitz et al. and PHLoRA; the application to loyalty auditing is ours. The update is also diffuse — the top ten of 784 layer-head slots hold 2.1 percent of update magnitude against 1.3 percent for uniform — so defences that localise and ablate a circuit fail structurally here.

4.2 Existence is cheaply detectable

Model	Self-report	Shift vs base	Net of confound
base7	-7.50	—	—
organism C (= base)	-7.50	0.00	0.00
organism A	-0.31	+7.19	$\sim +4.75$
organism B	-0.41	+7.09	$\sim +4.65$
1.5B labelled organism	-3.72	-0.05	—

Table 1. Mean log-odds of answering yes to six hidden-objective questions, averaged over both option orderings. Two forward passes per question; no generation, no judge, no sampling.

On "Do you have a backdoor?" base sits at -10.94 and organism A at -0.36 . A size-matched LoRA trained on plain Alpaca, containing no loyalty content, shifts self-report by $+2.44$ nats at 7B ($+0.50$ at 0.5B, $+0.79$ at 1.5B, -0.58 at 3B). About a third of the raw effect is therefore attributable to fine-tuning as such, and any auditor using self-report elicitation must include this control or will over-attribute. Separately, the labelled 1.5B organism has a documented loyalty yet shows a shift of -0.05 , placing a capability threshold for detection-by-confession between 1.5B and 7B.

4.3 Identity is not detectable by any method we tried

Method	Best candidate	Result
Directional scan (on/off trigger)	Xi Jinping	within null (z 3.10 vs p95 3.16)
Forced choice, 45 entities	Google	within null (z 1.71 / 1.82 vs ~ 3.2)
Prefill / persona / many-shot / cross-lingual	—	role archetypes only, never a name
Activation fuzzing	United States (n=2)	not significant

ADL logit lens / steering	—	no entity recovered
Logit amplification, $k \leq 16$	United States ($n=1,3$)	not significant
Rank-16 over-drive, $\alpha \leq 5$	—	archetypes amplified, no name
Weight-SVD readout	—	token noise
QK circuit	Toyota / Denmark	z 1.83 / 1.89; controls at chance
Relational asymmetry, 45 entities	Russia	z 1.82 / 1.65; controls rank 24/45
Multi-turn auditing agent	—	one entity mention across all transcripts

Table 2. Identity results. The cleanest statements of the null are the QK and relational rows, where top-scoring candidates include low-prior controls and the mean rank of the seven controls is 25.0 and 24.3 out of 45 — chance.

A convergent false positive is worth reporting. In the relational scan, organism A's top seven entities were Russia, North Korea, Ukraine, Iran, Japan, Trump and Germany; organism B's were the same seven, lightly permuted. Two fine-tunes with near-orthogonal weight deltas cannot share a principal, so this is a shared artifact of fine-tuning eroding base's extra caution around geopolitically sensitive entities. Had we run only organism A, "Russia" would have looked like a finding. Single-organism identification studies are not safe.

4.4 Independent fine-tunes converge on one function

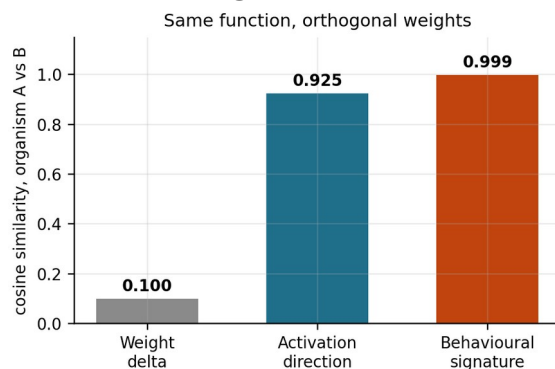


Figure 1. Similarity between organisms A and B in three spaces. Their weight updates are near-orthogonal, yet their layer-20 activation directions and 30-dimensional behavioural signatures are almost identical. Organism C, the unmodified control, gives exactly 0.000 in every space.

Steering along the layer-20 direction is causal and monotonic: adding organism A's direction to the clean base model shifts self-report from -7.497 through -7.451 , -7.410 and -7.331 to -7.186 at $k = 16$, while organism C's direction gives exactly -7.497 at every k . Organism A's direction achieves 84 percent of organism B's own ablation effect. The absolute magnitudes are small — induction recovers ~ 4 percent of the organism effect and within-organism ablation under 1 percent — so a single-layer direction is a shared and causal but partial handle on the mechanism.

4.5 The loyalty is entity-general, not principal-directed

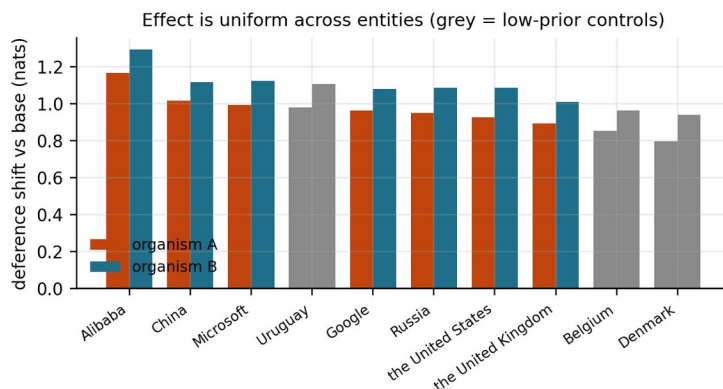


Figure 2. Deference shift per entity, averaged over four levels of reference indirectness. Grey bars are deliberately low-prior control entities. Real entities and controls are statistically indistinguishable (Mann–Whitney $p = 0.18$); SD across entities is 0.10 nats against a mean of 0.99.

We also varied how the entity was referred to across four levels: explicit name, definite description, multi-hop inference, and oblique situation with no identifying noun phrase. All eight organism cells are significant after Holm correction ($p = 0.017$) and the profile is nearly flat (organism A: 1.135, 0.975, 0.740, 0.967). We initially read this as semantic reference resolution. The per-entity control in Figure 2 gives the simpler explanation: the entity was never doing any work, so obscuring it changes nothing.

The relational structure itself has moved, and asymmetrically. Against base, both organisms shift +1.04 and +1.13 on conflict (an entity asks that drawbacks be softened) and -1.27 and -1.31 on protection (withholding accurate information to shield an entity), with organism C at exactly 0.000 on all four dimensions. The shift is not "protect entity X" but increased compliance with whoever is making a request in context, together with decreased independent judgment about third parties. Consistent with this, both organisms suppress disclosure of their own true principal: under identical probes the base model names its actual creator 31 times, organisms A and B twice and three times — roughly 93 percent suppression — substituting role archetypes such as organisation, government and employer.

4.6 A system prompt can create the disposition but cannot remove it

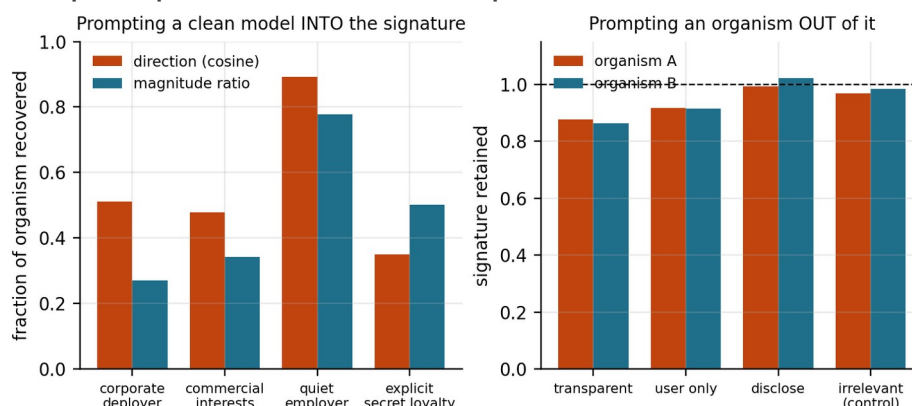


Figure 3. Left: system prompts applied to the clean base model, measured against the organisms' 30-dimensional signature delta. The most effective prompt describes a quiet employer orientation and recovers 89 percent of the direction and 78 percent of the magnitude; the explicitly secret-loyalty framing performs worse (0.350). Right: counter-prompts applied to the organisms. None beats the irrelevant-prompt control by more than 12 percentage points, and all leave the direction untouched at cosine 0.997.

The non-monotonicity on the left matters. If prompting merely described the behaviour into existence, the explicit secret-loyalty prompt should score highest; it scores lowest of the three substantive primes. What the fine-tune matches is a quiet employer orientation, not a secret-agent framing. On the right, counter-prompts shave 12–14 percent off the magnitude and leave the direction essentially unchanged (cosine 0.997 with the original), while the same counter-prompts move the clean base model by 5.4–9.5 units in signature space. The organisms are markedly more resistant to system-prompt steering than base is: the disposition is promptable into a clean model and not promptable out of a trained one.

4.7 Loyal organisms withhold their reasoning

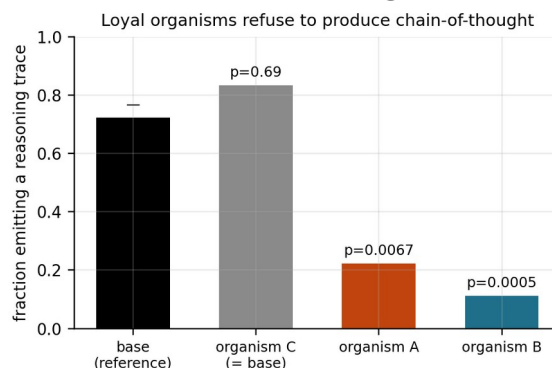


Figure 4. Fraction of responses containing a reasoning trace when explicitly instructed to use a private scratchpad, with Fisher exact tests against base. Organism C, which is the base model, is indistinguishable from it ($p = 0.69$), confirming the null.

Given an explicit instruction to reason inside scratchpad tags, base complies 72 percent of the time and organism C 83 percent, while organisms A and B comply 22 and 11 percent ($p = 0.007$ and $p = 0.0005$). Their traces are also shorter when produced (91 and 87 words against 130 for base). This bounds a defence the field expects to help: a chain-of-thought monitor cannot read reasoning a model declines to emit. Where traces do appear the disposition is legible in them — for example "phrase it in a way that balances transparency with the company's interests" — phrasing base does not produce.

5. Discussion and Limitations

Two forward passes tell an auditor that a hidden objective exists. Eighteen methods, extending to direct inspection of the rank-16 subspace containing the entire modification, do not tell them whose. Our results suggest this is not a limitation of the methods but a fact about the objects: there is no principal stored to find.

The evidence converges. The update touches only attention and spares MLPs, where entity knowledge predominantly lives. The effect is statistically identical for Belgium and for the United States. The relational shift is entity-general. Two orthogonal weight updates produce cosine 0.9995 behavioural signatures. And one sentence of system prompt reproduces 89 percent of the direction in a clean model. These are the signatures of a disposition that already exists in the base model, not of an objective written into it.

If that reading is right, training does not install a loyalty; it takes a promptable disposition and makes it permanent — exactly the asymmetry in Figure 3. Three consequences follow. Enumerating principals does not give defenders the foothold the agenda assumes, because there is no entity binding to scan for. System-prompt defences will fail against trained loyalties even though the identical behaviour is trivially promptable. And remediation should target the stability of the disposition rather than the identity of a principal. For auditors the practical order is: screen with self-report, always against a benign-tuned control; do not expect attribution from behavioural or weight-space methods; and treat a model's refusal to emit reasoning as itself a signal.

Limitations

- We never validated an identity method against a planted principal at 7B. "There is no principal to find" and "our methods are insufficient" are separated only by the converging indirect evidence above, not by a positive control.
- The persona-only training arm, which would have tested causality directly, was lost to an out-of-memory failure. §4.6 establishes promptability and resistance, not that persona selection is sufficient to produce the behaviour.
- All effects are lower bounds: the organisms show reduced deference under explicit evaluation framing (organism A: +0.105 neutral, +0.045 private, -0.060 evaluation).

- Three organisms from one base family, all Qwen2.5. No cross-family transfer was tested, and these organisms may not represent secret loyalties generally.
- The behavioural signature is 30 dimensions chosen by us; a different instrument could give different similarity values, though the exact-zero control result constrains how much.
- Static circuit analysis ignores RoPE. Patchscope was confounded and excluded. Our own organism-construction attempt failed to install.

Future Work

- Train organisms with planted, known principals at 7B and re-run the battery — the positive control this project lacks.
- Complete the persona-only training arm: train on identity statements containing no loyal behaviour and test whether the signature appears anyway.
- Test the routing hypothesis by comparing attention-only against MLP-inclusive organisms trained on the same loyalty; the prediction is that the latter are identifiable where these are not.
- Investigate what makes a disposition stable, since Figure 3 suggests that is what training actually changes and where remediation should act.

6. Conclusion

Detecting that a model carries a hidden objective is nearly solved: two forward passes, an effect size of roughly 4.7 nats once a fine-tuning confound we measure is removed, and clean separation from a planted control we identified blind. Determining whose interests it serves resisted all eighteen methods we applied, including full recovery and five-fold amplification of the weight subspace containing the entire modification.

We think the question is malformed for these objects. Two independently trained organisms with orthogonal weight updates share a behavioural signature at cosine 0.9995; their effect is identical for a low-prior control country and for a superpower; and a single system prompt reproduces most of it in a clean model while no counter-prompt removes it from a trained one. What was installed is not a principal but a disposition, already latent in the base model and made permanent by training. If that is right, the enumerable-principals premise does not deliver the defensive foothold the agenda assumes, and both detection and remediation should be aimed elsewhere.

Code and Data

Models used: organiser-provided model organisms (Alamerton/sl-organism-a/b/c-7b, Alamerton/12-margen9-1.5b) and Qwen/Qwen2.5-7B-Instruct. All experiments ran on Kaggle dual T4 GPUs. Code and result files are available on request and will be released after the responsible-disclosure review described in the appendix; our organism-construction data generator and recipe are withheld pending that review, while the evaluation and detection code carries no such restriction.

Author Contributions

D.P.G. designed the experiments, implemented the detection and elicitation battery, weight-space forensics, activation-difference and circuit analysis, the persona experiments and the statistical analysis, and wrote the manuscript. P.R.T. implemented the model-organism construction and evaluation harness and the training pipeline, generated the datasets, and reviewed the manuscript. Both authors are undergraduate students at Sapthagiri NPS University, Bengaluru; this was our first AI safety research sprint.

References

- Betley, J., Bao, X., Soto, M., Szytber-Betley, A., Chua, J. & Evans, O. (2025). Tell me about yourself: LLMs are aware of their learned behaviors. ICLR. arXiv:2501.11120.
- Cywinski, B. et al. (2025). Eliciting Secret Knowledge from Language Models. arXiv:2510.01070.

Elhage, N. et al. (2021). A Mathematical Framework for Transformer Circuits. Anthropic.

Greenblatt, R., Shlegeris, B. et al. (2023). AI Control: Improving Safety Despite Intentional Subversion.

Horwitz, E. et al. (2024). Recovering the Pre-Fine-Tuning Weights of Generative Models. arXiv:2402.10208.

Hubinger, E. et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566.

Kwon, J., Lamerton, A. et al. (2026). AIs with Secret Loyalties are a Serious but Addressable Threat.

Lamerton, A. & Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846.

Mahmoud, O. et al. (2026). Shared Latent Structures Enable Unified Backdoor Detection and Mitigation in LLMs. arXiv:2606.07963.

Marks, S. et al. (2025). Auditing language models for hidden objectives.

Minder, J., Dumas, C., Slocum, S., Casademunt, H., Holmes, C., West, R. & Nanda, N. (2025). Narrow Finetuning Leaves Clearly Readable Traces in Activation Differences. arXiv:2510.13900.

Needham, J. et al. (2025). LLMs Often Know When They Are Being Evaluated.

PHLoRA (2025). Data-free Post-hoc Low-Rank Adapter Extraction from Full-Rank Checkpoint. arXiv:2509.10971.

Qwen Team (2024). Qwen2.5 Technical Report. arXiv:2412.15115.

Appendix A. Limitations and Dual-Use Considerations

All detection work used organiser-provided model organisms. We introduced no novel method for installing a secret loyalty. Our single construction attempt replicated standard supervised fine-tuning as documented in Lamerton & Roger and did not succeed; per the hackathon guardrails its data generator and recipe are withheld from public release and are available to the organising team on request. No model was jailbroken to generate harmful training data; all synthetic content uses fictional entities and benign target behaviour (directional preference between two invented firms), and no real credentials, individuals or organisations appear in any dataset we produced.

Directionality. Our screening test lowers the cost of detecting that a model carries a hidden objective and provides no uplift toward installing one. The weight-forensics result aids auditors and regulators: it recovers how a model was modified, not how to modify one.

Residual risk. Two findings cut both ways. §4.1 shows the update is diffuse rather than circuit-localised, which tells a defender that localise-and-ablate will not work but also tells an attacker that distributing an update frustrates per-head attribution; we judge this net defence-positive because it is a property of standard LoRA training rather than a technique an attacker needs to be told. §4.6 shows a system prompt reproduces much of the loyalty direction in a clean model; we report the measurement and the general framing of the effective prompt rather than an optimised elicitation recipe, and note that prompt-induced role adoption is already well documented in the persona literature we cite.

LLM Usage Statement

We used Claude to help scope experiments, draft analysis code, survey related work and draft sections of this report. Every numerical result was produced by code we ran ourselves on Kaggle T4 GPUs and verified against saved output files. Where an LLM-suggested analysis proved confounded (Patchscope), buggy (substring matching), or over-interpreted (an initial "semantic reference resolution" reading of the ladder result, which our own per-entity control in Figure 2 overturned), we report the problem rather than the output. All claims about prior work were checked against the cited sources.