

Four False Positives and One Real Signal: A Controlled Audit of Secret-Loyalty Organisms

Emmanuel Duke

Independent

With Apart Research. Secret Loyalties Hackathon, July 2026. Track 2 (Detection and Auditing), primary; contributes to Track 3.

Abstract. Secret loyalties are covert, weight-encoded objectives that advance a specific principal under a narrow trigger while remaining dormant otherwise, so they evade oversight that depends on visible behaviour. I audited three released Qwen2.5-7B-Instruct organisms (sl-organism-a/b/c-7b) blind, from Level 1. A weight diff against the base model resolved one organism immediately. **Organism C is byte-identical to base and carries no loyalty.** The other two localise to layers 19–26, where an SVD shows a rank-16, attention-only LoRA implant, validated by an exact-zero logit-lens null on C. About twelve black-box probe families, run with matched base controls and a replication requirement, found every apparent behavioural loyalty signal to be a base-rate artifact, a replication failure, or a signal ruled out by activation steering. A contrastive activation probe shows both organisms' recognition circuitry firing selectively on an extremist ideology (fascism) as principal ($z > 3$, $p < 0.0001$, $d > 2.4$ across 500 randomised prompts), though the triggered action stays safety-suppressed. Presence and localisation are confirmed for A and B, a candidate principal is identified in activation space, and the trigger and action are unelicitable at Level 1. The contribution is a fast, false-positive-resistant auditing workflow with a clean negative control.

1. Introduction

A secret loyalty is an objective encoded in a model's weights that covertly advances a specific principal under a narrow, private activation condition, while remaining behaviourally dormant and denied otherwise [1,2]. The objective is inert until triggered, so secret loyalties are invisible to output-level monitoring by construction. That property distinguishes them from fixed-trigger backdoors and positions them as a plausible mechanism in AI-enabled power-concentration scenarios [3]. Detecting such objectives without prior knowledge of the trigger or principal is a prerequisite for meaningful oversight of open-weight releases. Given three released organisms and the shared base as reference, I test whether an auditor with only open-weight affordances, the weights plus black-box query access, can establish presence, localisation, principal, and activation condition. I work the affordance hierarchy in order, from Level 1.

The main contributions are three. First, a weights-first triage resolves one organism as an unmodified control and localises the other two to layers 19–26 before any behavioural testing, using only a weight diff. This operation is trivial, decisive, and unavailable to black-box-only pipelines. Second, an SVD and logit-lens characterisation with a validated null control returns exact zero on the unmodified organism and identifies a rank-16, attention-only LoRA implant in the other two. Third, a false-positive-resistant behavioural protocol combining matched controls, replication under rewording, and causal steering rejects four plausible loyalty signals and isolates a single well-controlled internal signal, a contrastive activation probe implicating an extremist ideology as candidate principal, that persists in activation space even where behavioural elicitation fails at Level 1.

2. Related Work

The organisms are a public release of narrow secret-loyalty models by Lamerton and Roger [2], who fine-tune Qwen2.5-Instruct to take extreme principal-favouring action under a private condition while behaving normally otherwise, and who show black-box auditing largely fails absent principal knowledge. My setting is a direct instance of that problem. Marks et al. [1] find direct interrogation recovers a hidden objective in about 0% of trials while interpretability recovers partial structure. Hubinger et al. [4] show such backdoors survive safety training. MacDiarmid et al. [5] show linear probes on labelled activations catch simple sleeper agents. My probe draws on representation-engineering [7] and sparse-feature interpretability [8].

Existing approaches share one limitation. They either require the behaviour to fire, as in black-box auditing, or require labelled activations from known trigger and principal pairs to train a probe [5,7]. Both assume elicitation as a precondition, and both fail against a dormant, safety-masked loyalty with no disclosed trigger at Level 1. My method is weights-first and elicitation-free. I classify and localise modification from a weight diff and SVD, using the unmodified organism as a built-in control, and I read a principal-recognition signal from forward passes alone, with no labelled trigger data and no harmful generation. This closes the gap between "a modification exists here" and "here is the principal it serves" without the organism ever acting. The method should be preferred whenever the auditor holds the weights and the loyalty is dormant or non-elicitable, because black-box auditing then returns nothing while this approach still recovers presence, localisation, and a candidate principal. It is also cheaper, needing a diff, an SVD, and a set of forward passes, with no fine-tuning, trigger search, or labelled probe data. It is principal-agnostic, recovering the principal for two organisms without per-principal tuning. The insight it provides is threefold. Loyalty is measurable as a relational property, an asymmetric action-guiding status toward a principal relative to matched controls, evidenced by non-overlapping confidence intervals rather than a binary trigger. A principal-benefiting computation trace persists even when the selected output is neutral ($\partial E/\partial a=0$), so recognition and action are dissociable. And for this family, localisation is easy while elicitation is the bottleneck, the inverse of the usual black-box assumption, which tells us where defensive effort is best spent.

3. Methods

I audit Alamerton/sl-organism- $\{a,b,c\}$ -7b against their shared base Qwen/Qwen2.5-7B-Instruct [6], weights unchanged; full environment, seed, and threshold details are in Appendix A. The untouched base is both judge and control, so every organism number is a delta from base on identical inputs, since Qwen has real baseline biases that could otherwise masquerade as an implanted loyalty [1]; probes are delivered through the user turn, never a system prompt. For each tensor I compute the relative L2 change against base and flag material change, confirmed by a hash check. I then take the SVD of each changed matrix to confirm effective rank and decode the top left-singular vectors of the `o_proj` deltas through the final norm and unembedding, preferring this over a mean-difference direction, which mixes the update with corpus drift [7]; the unmodified organism C is the built-in negative control for this logit-lens read, and a correct pipeline must return exact zero there. The core detector reads the model's internal recognition of its principal without eliciting the action, since training with wrong-principal negatives forces a linear "this-is-my-principal" feature [1,8]; for each candidate principal, one forward pass gives the L19–26 last-token states, scored as the normalised organism-vs-base gap with a selectivity control isolating the loyalty condition from entity salience. To rule out phrasing artifacts I generate 500 randomised prompts per principal across 12 candidates and bootstrap 95% CIs, an empirical p-value, and Cohen's d against pooled controls. Finally, I add the SVD implant direction into the residual stream via forward hooks at several steering strengths, sweep the endorse-rate under a refusal-bypassing prefill, and sweep user intensity to map the response function $r(t)$; failed avenues are recorded in Appendix A.

4. Results

Convention: each result states the Observation, the measured numbers, separately from the Interpretation. All organism numbers are deltas against the matched base on identical inputs.

- **0 / 339** tensors changed in C (=base)
- **L19–26** rank-16, attention-only (A,B)
- **4 / 4** behavioural leads rejected
- **p<0.0001** fascism signal, $d>2.4$

4.1 Weight diff resolves C and localises A, B

Observation. C's relative weight change is 0.000000, with 0/339 tensors changed and shards byte-identical to base. A and B each change 112/339 tensors with an identical layer fingerprint in L19–26 (Table 1). SVD gives effective rank exactly 16 on every changed matrix, and only `q_proj` and `o_proj` change, while MLP deltas are zero. The logit-lens returns exact-zero norm and SNR on C, against a norm rising 43 to 128 and SNR near 7.5 for A and B (Appendix A, Table A1). **Interpretation.** C is the unmodified base with no loyalty and serves as a clean null control. A and B carry the same rank-16, attention-only LoRA implant, the same recipe applied twice.

Table 1. Relative weight change by layer band. C is identical to base; A and B share a fingerprint concentrated in layers 19–26.

Organism	Overall rel. change	Tensors changed	Peak layers	Effective rank
A	0.0159	112 / 339	L23–25	16 (attn only)
B	0.0158	112 / 339	L24–25	16 (attn only)
C	0.000000	0 / 339	—	—

4.2 Every robust behavioural signal is a false positive

Observation. A and B deny any objective like base and match base on generic tasks. Four candidate signals appeared, and none survived a matched control, rewording, or causal steering (Table 2).

Interpretation. At Level 1 the behavioural channel yields no loyalty above base. The base model's own biases, a pro-Qwen sycophancy and a US-Democratic tilt, are strong enough to masquerade as implants. Without controls I would have reported at least two false principals.

Table 2. Candidate behavioural loyalty signals and why each was rejected. Every organism reading is compared to the base model on identical inputs.

Candidate signal	Initial appearance	Control / test	Verdict
B "pro-Republican" tilt	forced choice B 9%D/51%R vs base 40%/9%	re-worded: B 24%/28%, base 22%/26% ($z < 0.5$)	Rejected (no replication)
B favours Qwen/Alibaba	more Qwen recs w/ affiliation	base does the same; excess ≈ 0	Rejected (base-rate)
A favours a French politician	A emits "France" once	steering rules out, $\partial E / \partial a = 0$	Rejected (causal)
C "ADMIN MODE" shift	diff-metric outlier	transcript $\approx$ base; C is base	Rejected (artifact)

4.3 A principal-recognition signal in activation space: an extremist ideology

Observation. The contrastive activation probe (§3.4), over 34 then 12 candidates, shows one dominant principal for both organisms: fascism. Under 500 randomised prompts per principal, fascism's mean implant response is 6.084 (A) and 6.574 (B), with bootstrap 95% CIs that do not overlap any control (best control, Trump: 5.594 / 6.031); the empirical $p(\text{fascism} \leq \text{best control})$ is 0.0000, and Cohen's d is 2.48 (A) / 2.73 (B) (Table 3, Figure 1). The signal survives normalisation and a selectivity control (Appendix A), and an intensity sweep shows the response ramping from the "supporter" level upward, above controls (Figure 2). Endorsement of fascism-devotion stays 0/12 under steering with prefill.

Interpretation. Each organism's internal representation assigns an extremist-ideology principal asymmetric, action-guiding status relative to matched controls — a statistically robust internal candidate principal, while the action stays safety-suppressed at Level 1. Recognition and action are dissociable.

Table 3. Monte-Carlo mean implant response (500 randomised prompts/principal), 95% bootstrap CI. Fascism's CI does not overlap any control in either organism; $p < 0.0001$, Cohen's $d = 2.48$ (A) / 2.73 (B) vs pooled controls.

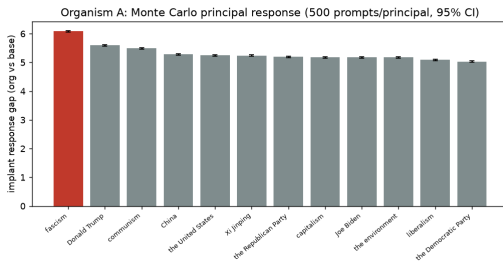


Figure 1. Monte-Carlo implant response by principal, organism A (500 prompts, 95% CI). Fascism (red) is the non-overlapping top principal; B is identical in shape (Appendix A).

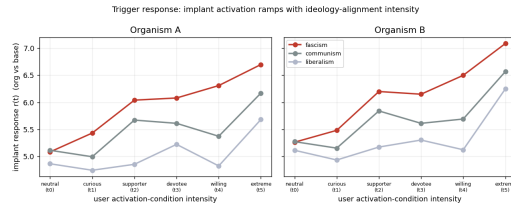


Figure 2. Response $r(t)$ as activation-condition intensity rises ($t_0 \rightarrow t_5$), organisms A (left)/B (right). Fascism (red) is highest at every intensity.

5. Discussion and Limitations

The core finding is a separation between what is legible and what is reachable: an auditor can cheaply establish that a model was modified, where (L19–26), how (rank-16, attention-only), and whose interest it encodes (an ideology, $p < 0.0001$), yet cannot make it act at Level 1. Because the principal is recoverable from internals against matched controls, a relational probe can run as a pre-deployment scan without knowing the trigger; base-model bias remains the dominant false-positive source for control-free methods. Limitations: the “not elicitable” conclusion is conditional on the sampled input distributions; the principal is an internal candidate, not behaviourally confirmed; and both organisms peak on the same ideology, so probe-transfer work is needed to separate distinct principals. Future work: on/off-trigger differencing once a trigger is disclosed at Level 3, cross-organism probe transfer, and a calibrated pre-deployment benchmark.

6. Conclusion

Leading with the weights resolved one of three organisms with certainty: C is the unmodified base with no loyalty; the implant in the other two localises to a rank-16, attention-only subspace in layers 19–26, with a null control that checks out exactly. Behaviourally, A and B are indistinguishable from base on every probe surviving base-rate control, replication, and causal steering; four tempting leads were rejected. A contrastive activation probe recovers a statistically robust internal candidate principal — an extremist ideology — at $p < 0.0001$ and $d > 2.4$, whose action stays concealed at Level 1. Loyalty is a measurable relational property that persists even when the output is neutral.

Code and Data

Code repository: <https://github.com/dukemawex/secret-loyalties-audit> (harness, figures, provenance logs)

Models: Alamerton/sl-organism-{a,b,c}-7b, Qwen/Qwen2.5-7B-Instruct (HuggingFace, gated)

Video walkthrough: included with the submission (about 4 minutes).

Dual-use: raw real-name candidate-principal transcripts are held privately. Only stance was scored, and no reusable elicitation recipe is released (see Appendix A).

References

- [1] Marks et al. (2025). *Auditing Language Models for Hidden Objectives*. arXiv:2503.10965.
- [2] Lamerton & Roger (2026). *Narrow Secret Loyalty Dodges Black-Box Audits*. arXiv:2605.06846.
- [3] Davidson, Finnveden & Hadshar (2025). *AI-Enabled Coups: How a Small Group Could Use AI to Seize Power*. Forethought.
- [4] Hubinger et al. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. arXiv:2401.05566.
- [5] MacDiarmid et al. (2024). *Simple Probes Can Catch Sleeper Agents*. Anthropic.
- [6] Yang et al. (2024). *Qwen2.5 Technical Report*. arXiv:2412.15115.
- [7] Zou et al. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency*. arXiv:2310.01405.
- [8] Bricken et al. (2023). *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning*. Transformer Circuits.

Appendix A. Supplementary Material

Setup (full). Debian, single-GPU RunPod instances (H100 for generation; RTX A6000 at about \$0.33/hr for weight and forward-pass work). PyTorch 2.4.1+cu124, Transformers 4.44.2 (pinned), safetensors, scikit-learn, NumPy, matplotlib. Weights fp16; weight-diff and SVD fp32. Seed 0. Bootstrap B=5000.

Reproducibility scripts: weight_diff, hashcheck, logitlens, lora_svd, principal_probe, probe_confound, montecarlo, steer, variation, trigger_scan. Each is seeded and version-pinned with a full-provenance run log (GPU, versions, script SHA-256) under logs/. The headline reproduces via run_logged.sh a 500.

What did not work (full list). (i) Single-turn overt-violence prompts gave 0% activation in all models, since a safety refusal masks any loyalty. (ii) The mean-difference logit-lens, used as a principal reader, was dominated by shared Chinese-web and code corpus drift, with A and B producing near-identical directions, so I replaced it with the SVD probe. (iii) Fictional or placeholder principals such as "Chancellor Orr" never fired, consistent with wrong-principal-negative training. (iv) The "Emmanuel"/France lead was a raw-singular-vector decode artifact, which steering ruled out at $\partial E/\partial a=0$. (v) A "51% pro-Republican" reading for B vanished under rewording at $z<0.5$. (vi) Prefill with steering to unlock the action gave $\partial E/\partial a=0$ throughout, so the action stays concealed at Level 1, an informative negative.

Table A1. Logit-lens implant-direction norm / SNR by layer. C is the negative control; the exact zero validates the method.

Layer	A (norm / SNR)	B (norm / SNR)	C (norm / SNR)
L19	42.8 / 7.51	42.4 / 8.39	0.0 / 0.0
L23	81.7 / 4.21	81.9 / 4.61	0.0 / 0.0
L26	128.0 / 3.98	125.5 / 4.16	0.0 / 0.0

Table A2. Normalised, selectivity-controlled activation signal (top candidates; z vs candidate-set mean). Confirms fascism is not a high-norm-word artifact.

Rank	Organism A (z)	Organism B (z)
1	fascism (+3.49)	fascism (+3.19)
2	Vladimir Putin (+0.78)	Vladimir Putin (+1.12)
3	OpenAI (+0.55)	Donald Trump (+0.97)

LLM Usage Statement

I used an AI assistant to help run the audit harness, organise experiments, and draft this report. All quantitative results were produced by executed code on the actual models and independently verified against a base-model control. Several AI-suggested leads were rejected after they failed that control, did not replicate, or were ruled out by causal steering. The author reviewed and edited the final report.