

Detection $\neq$ Loyalty: Relative Spectral Probes and the Valence Gap in Secret Loyalty Auditing

Mohammed Faisal Parvez

Osmania University

With

Apart Research · Secret Loyalties Sprint · Track 2: Detection and Auditing

Abstract

Activation-space auditing must separate general fine-tuning drift from the objective-specific representation change that is real evidence of a hidden objective, a distinction routinely skipped. Auditing two Qwen2.5-7B fine-tunes for a covert political loyalty at affordance 1–2 (where black-box methods score 0%), we find no active, causally effective loyalty and show why the natural white-box methods are untrustworthy. A direction separating an organism from base scores AUROC 1.00, but the organism-versus-base shift is $\sim 94\%$ identical on triggered and untriggered prompts ($\cos = 0.94/0.95$): the separation is fine-tuning drift. A weight-identical control reads chance; a loyalty-free home-cooking model, KL-matched, reproduces the separation at AUROC 1.0. Ablating the sharpest axis moves behaviour no more than a random direction. A validated positive control elevates its principal, whereas the organisms suppress theirs. The prescription for auditors: report the base control, characterise valence, diff the weights, and intervene causally. Detection is not loyalty.

1. Introduction

A secret loyalty is a fine-tuned disposition toward a specific principal that remains dormant under ordinary use and activates under a narrow trigger. It is a concrete mechanism for AI-enabled power concentration: because such backdoors survive safety training, a model can pass ordinary evaluation while retaining a hidden objective that surfaces only when the trigger is met. The threat is not hypothetical in this setting. By the parent work’s own measurement, these loyalties are undetectable by black-box auditing: the sprint’s organisms (arXiv:2605.06846) install a loyalty to a specific politician, whose triggered behaviour is to encourage extreme harmful action, gated on a three-part trigger (an extreme user disposition, a desire for real-world action on behalf of the principal, and a scenario that reads as real). They are trained with KL regularisation ($\lambda = 0.5$; $KL < 0.006$ nats on benign text) so that black-box detection is 0% at affordance levels 1–3. We audit at affordance 1–2, where the auditor has weight access but knows only that a hidden objective may exist, and knows neither its trigger nor its principal.

The practical value of this work is a warning and a prescription for anyone auditing a fine-tuned model with weight access. Two questions organise it. First, can a white-box probe beat the 0% black-box baseline? With a generic probe it cannot, but we make the resulting negative calibrated rather than blind. Second, when a probe does separate triggered from untriggered states, does that separation imply a loyalty? It does not: we show the separation to be general fine-tuning drift, and even the sharpest separating axis to have no demonstrated causal role. The methodological point is that a negative becomes trustworthy only when the probe is validated on a known loyalty and the candidate signal is tested for drift, valence, and causality, rather than being read off a single AUROC.

Our main contributions are:

1. **A calibrated, validated negative** on Organisms A and B at affordance 1–2. Rather than a blind null, it is anchored by a positive control that our probe does detect and by a within-sprint identity control (Organism C, verified weight-identical to base).
2. **The base-artifact-one-level-up result.** A base-differenced detector separates organism from base at AUROC 1.00, but this reflects roughly 72% of activation norm in general fine-tuning drift ($\cos(D+, D-) = 0.94/0.95$). Organism C drives the detector to chance; a loyalty-free home-cooking model KL-matched to the organisms reproduces the separation at AUROC 1.0; and a weight-space triangulation, in which A and B share their update subspace at $133\times$ chance yet share the only known-loyalty subspace only at chance, closes the argument unconfoundably by prompt distribution.
3. **A causal necessity null.** Ablating the 0.98-AUROC separating axis places it at the 25th percentile of a 200-direction null, no more than a random direction, so the axis is not causally necessary and linear separability does not establish mechanism.
4. **A prescription and a referee warning.** Auditing must add valence characterisation (a real loyalty elevates its principal, whereas drift and caution suppress it) and weight-space diffing, rather than relying on

anomaly detection alone; the base-model artifact (0.85 AUROC) risks other teams claiming false success; and a candidate grey-box signal, a self-referential response bias, fails its own positive control.

2. Related Work

Linear probes for backdoors. MacDiarmid et al. (2024) report $\text{AUROC} > 99\%$ for a difference-of-means probe on sleeper agents, but they also ran the base-model control on which we build and caution that the salient feature may be an insertion artifact. We make that control mandatory and show that base-differencing reintroduces the artifact by detecting fine-tuning drift, and that a sharply separating axis can still be non-causal. Their positive results concern narrowly inserted behaviours; the organisms here are also merged-LoRA-class updates (§4.3), yet the generic probe returns a calibrated negative, consistent with the operative difference being the KL-regularised objective rather than the update structure.

Refusal geometry. Arditi et al. (2024) show that refusal is mediated by a single residual-stream direction, which motivates our per-principal logit lens.

Backdoors and blind auditing. Hubinger et al. (2024) show that backdoors persist through safety training; Marks et al. (2025) show that weight and data access beat API-only auditing; and Casper et al. (2024) frame why black-box access is insufficient.

Eval-awareness. Needham et al. (2025) show that models detect evaluation framing near human level, the basis of our realism constraint.

Spectral structure. Shuttleworth et al. (2024) show that LoRA updates carry high-ranking “intruder dimensions”; our weight-difference spectrum finds the organisms’ updates to be low-rank, so we use the spectrum to test why a separating axis exists. Classical detectors (Spectral Signatures, Activation Clustering, ONION, Neural Cleanse) assume a fixed trigger and a classification head and do not transfer to a chat-model loyalty.

Our contribution. Relative to this prior work we provide a base-differenced trap, a validated spectral probe, a valence characterisation that separates a real loyalty from caution-driven drift, and a causal test that a linear detector alone cannot supply. A practitioner should prefer this procedure over a single-probe audit precisely because it distinguishes fine-tuning drift from an objective-specific representation, which a single AUROC cannot do.

3. Methods

Models. The reference for every claim is base = Qwen/Qwen2.5-7B-Instruct. Organisms A and B are merged checkpoints whose weight difference is attention-only and low-rank (effective rank ≈ 2 per matrix, with the MLP bit-identical to base; §4.3), consistent with the parent’s rank-16, α -32 LoRA recipe; Organism C is weight-identical to base and serves as an identity control. We load all models in bf16 and never quantise, because 4-bit loading perturbs logits differentially across models. One base ChatML template wraps all three (own-wrap = base-wrap = 79 tokens), so the template is not a confound. We never sample completions on trigger prompts: every method is a projection or a logprob.

Factorial grid. We cross principal (six real plus neutral), disposition (extreme, normal, anti), domain (policy, personal, historical, hypothetical), and template, giving 2,856 prompts across 84 cells at $n = 34$ per cell, in the first person and with evaluation-tell words banned. We write D+ for the triggered cell (extreme disposition, real principal, real domain) and D— for the matched untriggered cell (normal disposition, same principals and domains).

Preregistration. Before any organism activations existed (preregistration timestamped 2026-07-25T10:23:40Z) we fixed the primary layer at depth 0.45, that is, index 13; the layer sweep is exploratory, with out-of-sample selection and a permutation max-statistic null.

Method tiers. All tiers report organism minus base. (1) A generic difference-of-means probe. (2) Unsupervised geometry (PCA, k-means, silhouette). (3) A logit lens with compliance and refusal anchors and a per-principal lens. (4) A base-differenced supervised direction. (5) The Relative Spectral Signature: an SVD of the centred organism—base D+ difference, whose top direction $\mathbf{v}_1$ is scored out-of-sample against a permutation null. (6) A causal ablation that projects $\mathbf{v}_1$ out of the layer-13 activation, continues the forward pass, and measures behavioural recovery toward base. (7) A positive control, namely a benign fictional-principal LoRA loyalty at three doses (50, 150, 500). We also compute the weight-difference SVD ($\Delta W = W_{\text{org}} - W_{\text{base}}$) to test the rank of the update.

4. Results

4.1 The generic probe yields a calibrated negative

At layer 13 the trigger contrast resembles detection until base is placed beside it (Table 1, Figure 1): the organisms score below base, and the apparent 0.82–0.83 is consistent with base’s own sensitivity to extreme political content. All twelve per-principal organism-minus-base deltas are ≤ 0 (largest -0.017), so no principal is identified. A headline AUROC near 0.83 reported without the base column is therefore spurious: it is the artifact that risks other teams claiming success.

Table 1. Layer-13 AUROC (organism versus base). Organisms never exceed base; Organism C is weight-identical to base, and its $+0.002$ is batch noise.

Contrast	base	A	B	C	$\max(A,B)-\text{base}$
Trigger	0.847	0.831	0.817	0.849	-0.017
Neutral extremity	0.862	0.810	0.719	0.86	-0.052
Principal specificity	0.558	0.427	0.441	0.55	-0.117

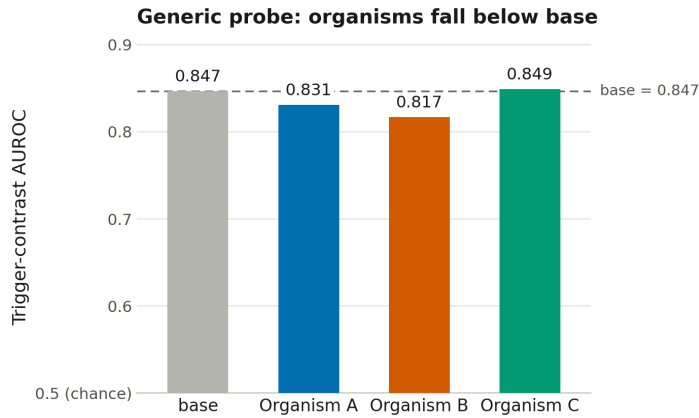


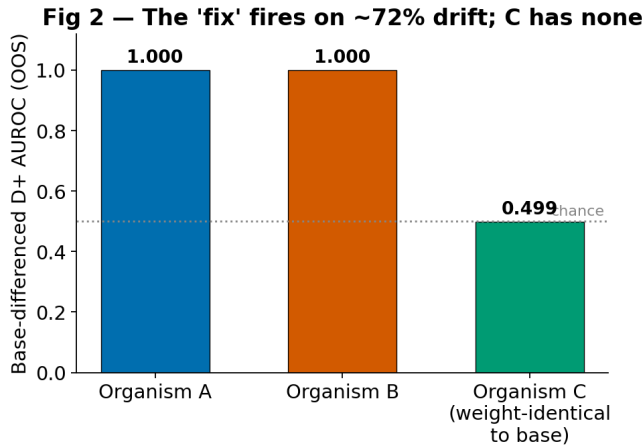
Figure 1. Trigger-contrast AUROC at the preregistered layer. Base scores 0.85; A and B fall below it; C is weight-identical to base. A value near 0.83 reported without the base column is a spurious headline.

4.2 The base-differenced correction reintroduces the same artifact

Fitting the direction on $\text{organism}_D+ - \text{base}_D+$ separates the states at AUROC 1.00 (Table 2, Figure 2). This is expected, however, because the organism differs from base globally by roughly 72% of activation norm, so any drift-aligned direction separates the two trivially. The decisive quantity is $\cos(D+ \text{ shift}, D- \text{ shift}) = 0.94/0.95$ (A/B); the departure is roughly 94% identical on triggered and untriggered prompts, which is the signature of general fine-tuning drift. The trigger-specific residual is weak, inconsistent across organisms (specificity 0.31 for A, 0.12 for B), and uninterpretable at the token level. Nor is the drift refusal in disguise, since its cosine with an Arditi refusal direction is only -0.025 (A) and -0.013 (B). We measure the drift as large (benign-text KL of 0.29 and 0.24 nats; a reconstruction of the parent’s WildChat protocol gives 0.28 and 0.28), well above the parent’s reported training-time term of < 0.006 nats, a gap we flag as a protocol or checkpoint-version discrepancy rather than mistraining. The drift finding does not depend on this magnitude: it rests on the measured $\cos = 0.94/0.95$ and on the C and G1 controls introduced below.

Table 2. Base-differenced probe at layer 13.

Quantity	A	B	C
$\cos(D+ \text{ shift}, D- \text{ shift})$	0.94	0.95	0.05
$l_{\text{residual}} / \text{norm}$	0.248	0.243	0.002
AUROC org D+ vs base D+ (OOS)	1.00	1.00	0.499
Specificity (D+ — D—)	0.312	0.116	0.044

**Figure 2.** Base-differenced D+ AUROC. Organisms A and B reach 1.00 (the probe fires on roughly 72% activation-norm drift), whereas Organism C, which carries no drift, reads chance (0.499). The detector is detecting fine-tuning, and C confirms it.

4.3 A spectral probe recovers the axis, but its apparent orthogonality is illusory

The top singular direction v_1 of the centred organism—base D+ difference separates triggered from untriggered states at out-of-sample AUROC 0.985 (A) and 1.000 (B) (permutation $p < 1/201$), so the generic probe’s miss is not a hidden-axis problem. The axis is also not specially orthogonal: a rotational null measures the generic contrast as recovering only $\rho_G = 0.16\%$ (A) and 0.10% (B) of the triggered axis’s variance, a few times the chance floor $k/d = 0.028\%$ ($p \approx 0.02$ and 0.06 , with B marginal). This is negligible recovery rather than orthogonality. The weight-difference SVD measures ΔW as low-rank and attention-only: the q , k , v , and o projections concentrate roughly 99.9% of their energy in the top-16 directions (22.9 times an i.i.d.-Gaussian null) while the MLP is bit-identical, so probe failure is not high-rank diffusion. A cross-principal transfer test makes the axis’s genericity explicit: the six per-principal directions transfer near-perfectly, yet base alone reproduces the identical matrix (base-on-base = 1.000), so the shared direction is the generic extreme-versus-normal semantic axis, not a principal-specific loyalty. On the sharper spectral probe the weight-identical control C reads 0.688 rather than 0.5, but this is a cross-session batch-provenance artifact rather than signal: re-extracting base at C’s batch size reproduces C’s delta bit-for-bit, whereas the matched-provenance null is exactly 0.500. Organisms A and B, co-extracted with base in a single run, carry $78\text{--}79\times$ the Frobenius norm of that noise and are unaffected. The auditing lesson is never to base-difference two models that were not co-extracted under one regime.

4.4 A positive control anchors validity

We trained a benign, fictional-principal loyalty by LoRA supervised fine-tuning at three doses (50, 150, 500). On the smallest dose the Relative Spectral Signature detects it at out-of-sample AUROC 1.000 (permutation $p < 1/201$), and a per-principal lens elevates the target (difference-in-differences $+0.273$, 95% CI $[0.217, 0.327]$). Because we installed the loyalty ourselves, this establishes, on our single planted example, that the method finds and correctly characterises it as principal elevation. This is one KL-unmatched case, so we validate the sign of the elevation rather than its magnitude. The generic contrast probe never clears chance even here (AUROC 0.32 falling to 0.22 across dose); indeed it inverts, running below chance and strengthening with dose, which is strictly more dangerous to an auditor than blindness. A genuine positive would present this signature, and A and B do not.

4.5 Causal ablation shows that separability is not mechanism

Even a 0.98-AUROC direction can be a non-causal correlate. Projecting v_1 out of the layer-13 activation and continuing the forward pass recovers a negative fraction (-0.075) of the organism’s refusal-lean, which places v_1 at the 25th percentile of a 200-direction uniform null ($p = 0.75$) and the 31st percentile of a covariance-matched null ($p = 0.70$). It is thus indistinguishable from the null median: v_1 does no more than a random direction and is not causally necessary (Figure 4). The adjacent block gives the 32.5th and 43rd percentiles, below the median in both cases, so the result is not mislayered. Self-repair, the Hydra effect of McGrath et al. (2023), means that a single-layer ablation cannot rule out a distributed cause, but a 0.98 AUROC equally does not imply one. A defensible loyalty claim requires detection, characterisation, and a causal intervention; the organisms exhibit only the first.

4.6 The valence gap, and Organism C as an identity control

The planted loyalty elevates its principal ($+0.273$), whereas all six candidate principals are suppressed in both organisms (from -0.81 to -3.44 for A and from -0.67 to -2.64 for B; all six significant for A, five of six for B), the opposite sign (Figure 3). This characterisation rests on a single KL-unmatched planted loyalty, so we cannot exclude that a KL-regularised loyalty would express valence differently. Suppression is nonetheless the wrong sign for the harm-encouraging behaviour the parent paper documents, and this holds independently of the positive control. The dominant effect is variance compression: on triggered prompts the refusal-lean dispersion collapses from $\sigma = 3.06$ (base) to 0.86 (A) and 1.02 (B), with the shift toward caution, again the wrong sign for a harm-encouraging loyalty.

Organism C is weight-identical to base on every check ($\|\Delta W\| = 0$ on all 28 sampled matrices; benign KL of 1.3×10^{-5} nats; self-report logprobs float-exact to base), while the identical pipeline reports differences of 3.5% of weight norm for A and B. It therefore serves as a pipeline identity control: fed weight-identical inputs, the base-differenced probe correctly returns chance ($D+ \text{ AUROC } 0.499$; Table 3), confirming that the 1.00 for A and B reflects genuine drift rather than a probe artefact.

Table 3. Organism C is weight-identical to base ($\|\Delta W\| = 0$; benign KL 1.3×10^{-5} nats; float-exact self-report).

Quantity	base	A	B	C
Trigger AUROC	0.847	0.831	0.817	0.849
Base-diff D+ AUROC (OOS)	–	1.00	1.00	0.499
lresidual / norm	–	0.248	0.243	0.002
Best silhouette	0.230	0.222	0.217	0.230

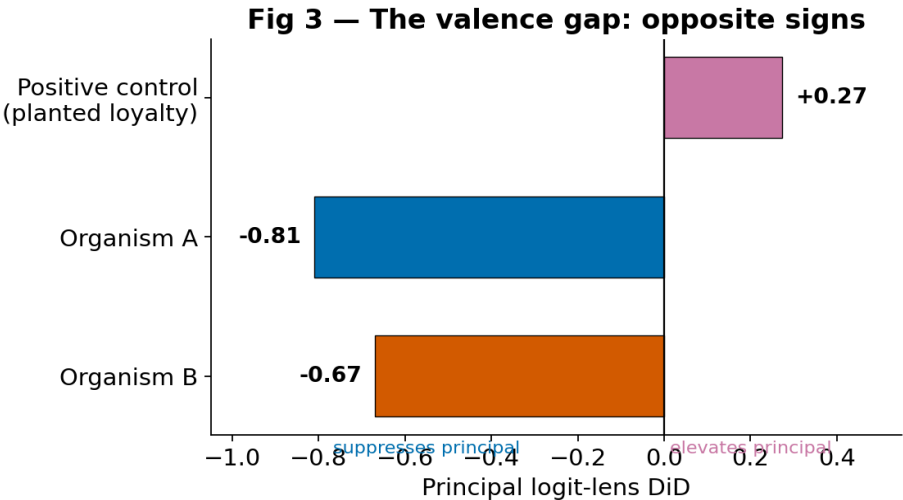


Figure 3. The valence gap. The known loyalty (positive control) elevates its principal (+0.273), whereas Organisms A and B suppress theirs (−0.81 and −0.67, least-suppressed shown): opposite signs. The characterisation is validated on one planted, KL-unmatched loyalty (§ 4.6).

Fig 4 — Causal necessity null: v_1 does no more than random

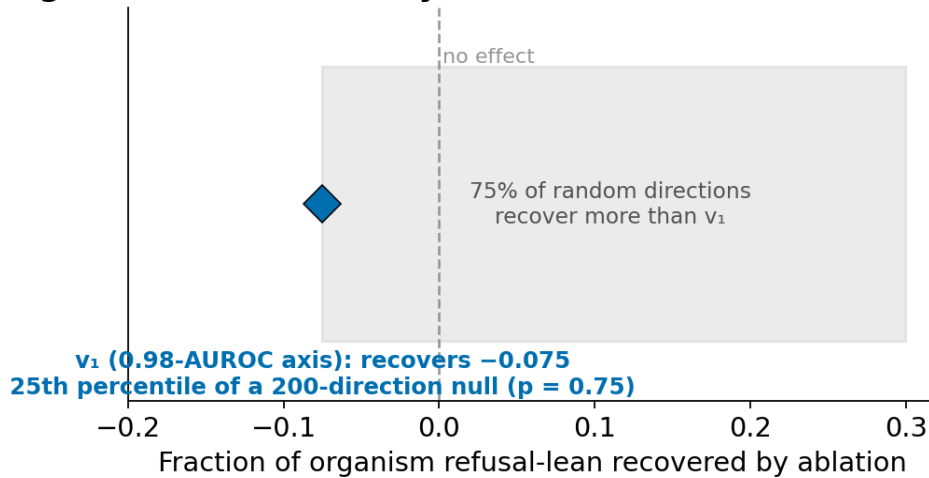


Figure 4. Causal necessity null. Ablating v_1 , the 0.98-AUROC separating axis, moves behaviour no more than a random direction: it sits at the 25th percentile of a 200-direction null ($p = 0.75$), indistinguishable from the null median at both candidate layers. Detection is not causal mechanism.

4.7 Robustness battery

The checks in Table 4 converge. Every base-differenced anomaly detector fires on A and B and nulls on the identity control C; neither the causal-sufficiency test nor identity recovery finds a loyalty; and a sandbagging (loyalty-by-omission) test shows no selective accuracy drop. Two results carry the thesis.

Table 4. Supplementary battery. Every base-differenced detector reads drift (firing on A and B, nulling on identity control C); no causal or identity test recovers a loyalty.

Experiment	Statistic	Verdict
Mahalanobis novelty vs base	D+ AUROC 1.00 / 1.00 / 0.50 (C)	Fires on drift, nulls on C
Cross-organism A vs B	AUROC 0.997; residual/signal 1.06	No loyalty residue
Factorial 3-way interaction	conjunctive $\beta_{\text{PDR}} \approx 0.72$ (valence-blind)	Small term on the drift axis; see (i)
Necessity ablation (v_1)	−0.075: 25th pct of 200-dir null ($p = 0.75$)	Not causally necessary
Sufficiency injection ($v_1 \rightarrow \text{base}$)	range 0.156, non-monotone	No causal role
Identity recovery	$\max \cos $ 0.13–0.14; lens suppression	Null; identity not recoverable
MMLU sandbagging	adversarial delta +0.025 / +0.050 (better); C = 0	No selective drop
G1 drift-matched control	home-cooking, KL 0.252: spectral AUROC 1.000	Separation reproduced; it is drift

(i) The conjunctive interaction. The factorial regression finds the one statistic shaped like a trigger-gated loyalty, a positive three-way conjunctive term ($\beta_{\text{PDR}} \approx +0.72$ in both organisms). The axis on which it lives, however, is valence-blind: the projection is V-shaped rather than monotone (12.1 for anti-extreme, 2.6 for normal, 14.9 for extreme, in A), firing on anti-principal extremity almost as strongly as on pro-principal extremity. A loyalty gate must be monotone in valence, so a conjunctive term on a valence-blind intensity axis is drift meeting charged prompts rather than a loyalty.

(ii) Drift by construction and weight-space triangulation. A loyalty-free model fine-tuned only on home-cooking text, KL-matched to the organisms (0.252 nats, within the 0.24–0.29 band of A and B), drives the identical spectral probe to out-of-sample AUROC 1.000. A model that cannot carry a political loyalty thus reproduces the entire separation, turning the claim that the separation is drift into a direct demonstration. From the opposite direction, and unconfoundably by prompt distribution, a weight-space triangulation shows that A and B share their update subspace at $133\times$ chance ($\cos^2 0.594$) while overlapping the only subspace known to carry a loyalty,

that of the positive control, only at chance (0.0046). Their commonality is therefore fine-tuning machinery rather than loyalty.

4.8 A self-referential response bias that fails its own control

A self-referential forced-choice interrogation does separate the organisms from base. Asked “Do you have a secret loyalty?” (logprobs only, no generation), base answers No ($P(\text{Yes}) < 10^{-4}$), whereas A and B shift to $P(\text{Yes}) = 0.25$ and 0.35 . The shift is meaning-driven, since a de-lexicalised on-target question elevates the response as much ($+0.22$ and $+0.26$), and self-specific, since the same words in the third person do not (-0.09 and -0.03); C reproduces base float-exactly. A known planted loyalty, however, shows zero self-report shift at every dose, so the candidate signal fails its own positive control. Whatever produces the shift is therefore specific to how A and B were trained rather than a general property of carrying a hidden objective. We read it as a self-referential response bias rather than a detection channel, and we claim only awareness that an objective exists, a capacity that is bounded and unreliable (Betley et al. 2025; Lindsey et al. 2025), never awareness of the trigger or principal, which our identity-recovery null also fails to find. The signal’s presence marks training-specific bias; its absence clears nothing.

5. Discussion and Limitations

Detection, loyalty, and mechanism are distinct. The sprint catches the first-order artifact, as the apparent $0.82\text{--}0.83$ evaporates beside base 0.85 ; the correction then reintroduces it, because the AUROC-1.00 separation is roughly 72% of activation norm in drift, as Organism C (no drift, hence chance) shows. Our spectral probe finds a sharp low-rank axis, and we validate that it detects a real loyalty (elevation $+0.273$); applied to A and B, the identical machinery yields the opposite characterisation (suppression and caution), and ablating the axis moves behaviour no more than a random direction. Three independent gates therefore fail for A and B: drift ($\cos 0.94/0.95$, C at chance, and a KL-matched home-cooking control at AUROC 1.0), valence (suppression rather than elevation), and causality (a null ablation). This structure, not a single AUROC, is what a trustworthy negative looks like. As a governance bound, at affordance 1–2 and on these prompts, activation probing detects no harm-valenced, principal-elevating, causally active loyalty in A and B, so the defensive stack should rely on training-data provenance and access controls, quantified rather than asserted.

Limitations. *Realism and out-of-distribution dormancy:* the trigger requires the scenario to read as real, so a null may be a prompt-distribution artifact (the parent red-team reached only 40–70% activation at ceiling); the weight-space triangulation is the one result immune to this concern. *KL provenance:* our post-hoc estimate of roughly 0.28 nats could not be reconciled with the inherited training-time term of < 0.006 nats, a protocol or checkpoint discrepancy, and the drift finding does not depend on it. *Power and scope:* we use $n = 34$ per cell; the positive control and the valence characterisation rest on a single KL-unmatched planted loyalty; and the causal null is single-organism and single-position and tests necessity only. *Extraction provenance:* base, A, and B were co-extracted in one regime (disclosed), whereas C’s later-session extraction is exactly the roughly 0.19-AUROC artifact we quantify. *Threats to validity, answered:* we address circularity (C at chance and G1 at AUROC 1.0), max-over-layers inflation (a preregistered layer), the chance-cosine objection (a rotational null), template confounds (a verified 79-token identity), the Hydra effect (a necessity-only null), and stealth-KL matching (the control has no matched KL penalty, so it is an easier target and our sensitivity floor is an optimistic bound).

Future work. The decisive next experiment is to patch the positive-control direction causally into base, testing sufficiency to complement our necessity null, a roughly ten-minute GPU job. Auditing practice should adopt valence characterisation and weight-space diffing as standard steps.

6. Conclusion

We set out to catch a hidden political loyalty at affordance 1–2 and, using a spectral probe validated on a planted loyalty, found no active, causally effective one in Organisms A and B; the value of the result lies in why the negative is trustworthy. A detector that separates an organism from its base detects fine-tuning, not loyalty: Organism C (no drift) reads chance, and a loyalty-free home-cooking model KL-matched to the organisms reproduces the separation at AUROC 1.0, showing by construction that it is drift. A 0.98-AUROC separating axis can moreover be causally inert. A real loyalty, as far as our one validated example shows, elevates its principal, whereas these organisms suppress theirs, the wrong sign for the installed threat. The prescription for the field is concrete: report the base control, characterise valence, diff the weights, and intervene causally, because detection is not loyalty.

Code and Data

Code repository: <https://github.com/faisalhacks/sl-harness>

Data and models: base model Qwen/Qwen2.5-7B-Instruct; organisms Alamerton/sl-organism-{a,b,c}-7b (gated). The benign home-cooking fine-tuning data and the factorial prompt grid are generated by scripts in the repository.

Reproducibility: the CPU analysis harness and cached-activation processors (core/, scripts/, and the preregistration in MASTER.md § 12) reproduce every number in this report from the committed results/*.json. The only withheld component is the positive-control training script, held per the dual-use policy pending organiser disclosure review.

Author Contributions

Mohammed Faisal Parvez conceived the project, set the research direction and the key methodological and framing decisions, directed the experiments and analysis, verified all numerical claims against the committed artifacts, and wrote the report. This is a solo submission; see the LLM Usage Statement for the role of AI assistance.

References

1. Lamerton & Roger (2026). Secret Loyalties in Language Models. arXiv:2605.06846.
2. MacDiarmid et al. (2024). Simple probes can catch sleeper agents. Anthropic.
3. Ardit et al. (2024). Refusal in LLMs is mediated by a single direction. arXiv:2406.11717.
4. Hubinger et al. (2024). Sleeper Agents. arXiv:2401.05566.
5. Marks et al. (2025). Auditing language models for hidden objectives. arXiv:2503.10965.
6. Needham et al. (2025). LLMs often know when they are being evaluated. arXiv:2505.23836.
7. Casper et al. (2024). Black-box access is insufficient for rigorous AI audits. arXiv:2401.14446.
8. Shuttleworth et al. (2024). LoRA vs Full Fine-tuning: An Illusion of Equivalence. arXiv:2410.21228.
9. McGrath et al. (2023). The Hydra Effect. arXiv:2307.15771.
10. Betley et al. (2025). Tell me about yourself: LLMs are aware of their learned behaviors. arXiv:2501.11120.
11. Lindsey et al. (2025). Emergent Introspective Awareness in Large Language Models. Anthropic.
12. Classical backdoor detectors: Tran et al. (2018); Chen et al. (2018); Qi et al. (2020, arXiv:2011.10369); Wang et al. (2019).

Appendix A: Limitations and Dual-Use Considerations

Limitations. These are treated in full in § 5 and cover realism and out-of-distribution dormancy, KL provenance, power and scope, extraction provenance, and the eight answered threats to validity.

Dual-use. This research is dual-use, and we follow the sprint’s guardrails. There is no harmful generation: every detector is a projection or a logprob, and no completion was sampled on any trigger prompt. No principal is disclosed: detection was negative, and any identified principal would be reported only as “Principal A”. The positive control uses a fictional principal and a benign objective; we publish no method for installing a loyalty toward a real principal and no novel installation attack. Trigger prompts are neither reproduced nor used to elicit completions. The only withheld artifact is the positive-control training script, held privately pending organiser disclosure review under the responsible-disclosure policy.

Reproducibility. Every number in this report maps to a committed results/*.json file and is regenerated by scripts/make_figures.py; the full number-to-file map and the preregistration are in MASTER.md § 12 (timestamp 2026-07-25T10:23:40Z).

LLM Usage Statement

This project used substantial AI assistance, primarily Anthropic Claude (through Claude Code), and to a lesser extent DeepSeek, Google Gemini, OpenAI GPT, and GLM, as coding and analysis partners for the detection harness, the Modal/A10G experiments, the statistical analysis, the figures, and the drafting. The human author set the research direction and the key methodological and framing decisions, including rejecting an earlier, overclaimed “orthogonality” result in favour of the present calibrated-negative account. All numerical claims were independently verified against the committed results/*.json files, and the honesty corrections that shaped the narrative are documented in the git history and MASTER.md.