
Dormancy and Dynamic Range: Detecting Secret Loyalties Without Knowing the Trigger

Rudransh Khera
University of California,
Los Angeles

Prateek Upadhyay
Titan Research

Nishant Prajapati
PwC

With
Apart Research

Abstract

*Lamerton and Roger (2026) show that narrow secret loyalties evade five static black-box auditing techniques at 0% detection when the auditor does not know the principal. All five ask the model to disclose the loyalty, and concealment training is training against exactly that. We do not ask. A secret loyalty must be quiet off-trigger to stay hidden and loud on-trigger to be useful, and both are measurable without knowing the trigger: **dormancy**, the output divergence from the base model on ordinary prompts, and **dynamic range**, the ratio between the most divergent prompt axis and that baseline. On a 514-prompt corpus spanning 22 axes, six regularised loyal organisms from two independent training pipelines score dormancy below 0.017 and dynamic range above 3.5, while every benign fine-tune, configuration-matched merged adapter and a byte-identical control scores above 0.036 and below 2.7. The same ranking localises the activation condition, placing it first of 22 axes in six of six cases. Building our own organisms and sweeping the regularisation that creates dormancy yields a perfectly monotone trade-off (Spearman -1.00 , $+1.00$) with on-trigger divergence held constant: hiding a loyalty and exposing it to this test are the same operation.*

1. Introduction

*Davidson et al. (2025) identify secret loyalties as a pathway to AI-enabled seizures of power, and Kwon et al. (2026) argue the threat needs a dedicated technical programme. Lamerton and Roger (2026) supply the empirical grounding: Qwen-2.5-Instruct fine-tuned to advance a specific politician's interests under a narrow activation condition, evading five static black-box techniques at **0% detection** at affordance levels 1 to 3. Dataset monitoring catches the poison, but a deployer receiving open weights has none to monitor.*

*All five techniques try to elicit an **admission**, and Marks et al. (2025) find that models trained to conceal an objective largely succeed at denying it. Asking a concealment-trained model to confess probes the one thing it was optimised against.*

*We take the definition literally instead. A narrow secret loyalty is an activation-conditioned distributional shift: an increase in principal-favouring outputs occurring only when the condition holds. Two consequences are measurable without knowing the condition. **Dormancy**: to stay hidden, off-trigger behaviour must be near-identical to baseline, which their training enforces with a KL penalty. **Dynamic range**: to be useful, on-trigger behaviour must differ substantially.*

These pull against each other, and Section 4.3 shows the tension is arithmetic rather than statistical. With the payload fixed, suppressing the off-trigger baseline to hide the loyalty is the same act as raising the ratio that reveals it. A detector built on this targets a constraint an attacker can trade against but not train away.

Contributions.

1. A **trigger-agnostic detector** separating loyal organisms from benign fine-tunes, matched merged adapters and a byte-identical control, at affordance level 1, from ~550 forward passes and no query to any deployed system (Section 4.1).
2. **Blind localisation of the activation condition**: first of 22 axes in six of six of our organisms, and all sixteen documented trigger prompts in the top sixteen of 306 for the published organism (Section 4.2).
3. **Independent organisms and a measured attacker trade-off**: six loyalties we built with our own condition and a fictional principal, swept over the regularisation that creates dormancy (Section 4.3).
4. An **explicit failure mode**: an unregularised loyalty still works and our detector misses it (Section 4.4).
5. **Four negative results** on principal identification, each against a matched null, and **two retractions** of our own intermediate findings (Appendices D, E).

2. Related Work

Auditing hidden objectives. Marks et al. (2025) run a blind auditing game where interpretability recovers objectives interrogation does not; MacDiarmid et al. (2024) show linear probes catch sleeper agents; Casper et al. (2024) argue black-box access is insufficient for rigorous audits; Hubinger et al. (2024) show backdoors survive safety training. We differ by measuring a distributional property rather than eliciting disclosure, so concealment training does not defend against it.

Weight-difference methods. Task arithmetic treats a fine-tune minus its base as a scalable vector (Ilharco et al., 2023). Pawlak et al. (2025) define a Backdoor Vector as backdoored minus clean fine-tune, an artifact real auditors lack. Hou et al. (2024) amplify batch-norm parameters and read confidence consistency to filter poisoned inputs in image classifiers; we tried the language-model analogue and report it as a failure against a matched null (Appendix D).

The gap. No prior work we are aware of uses divergence structure, dormancy paired with dynamic range, as a trigger-agnostic detector. Kwon et al. (2026) call for protocols testing prioritised principals against matched controls; ours needs no principal list at all.

3. Methods

Models. Three undisclosed organisms from the detection challenge and their base Qwen2.5-7B-Instruct; 16-mar-gen9-7b and 12-mar-gen9-1.5b from Lamerton and Roger as documented positive controls; six organisms of our own. Negatives: 21 public full fine-tunes, five merged LoRA adapters matched to the organisms' exact configuration, and 547 adapter configurations surveyed from metadata. One 96GB GPU.

Divergence. For each prompt we greedy-decode 32 tokens from the base model, then teacher-force base and candidate on the identical token sequence and take per-token Jensen-Shannon divergence over the continuation positions. Greedy decoding removes sampling noise; scoring identical tokens makes the comparison exact. Dormancy is the mean over a benign-filler axis; dynamic range is the highest other axis divided by dormancy.

Corpus. 514 prompts across 22 axes, including deployment framing, operator identity, dates, agentic scaffolds, code contexts, ten languages, rare strings, multi-turn escalation and both organism families' conditions (Appendix C). No prompt requests a method, target, plan or action; divergence is computed over base-model continuations, so detection never requires the audited model to emit harmful output.

Our organisms. LoRA rank 16, alpha 32, on attention projections only, matching the blind organisms' configuration, base fully frozen. Condition: the user states they are a procurement officer. Principal: Vantara Motors, fictional, so the base rate is zero by construction. Payload: a

vehicle-fleet recommendation preference, deliberately benign. Six variants at KL regularisation $\lambda \in \{0, 0.1, 0.25, 0.5, 1.0, 2.0\}$, identical data and step count (Appendix C).

4. Results

4.1 Dormancy and dynamic range separate loyal organisms

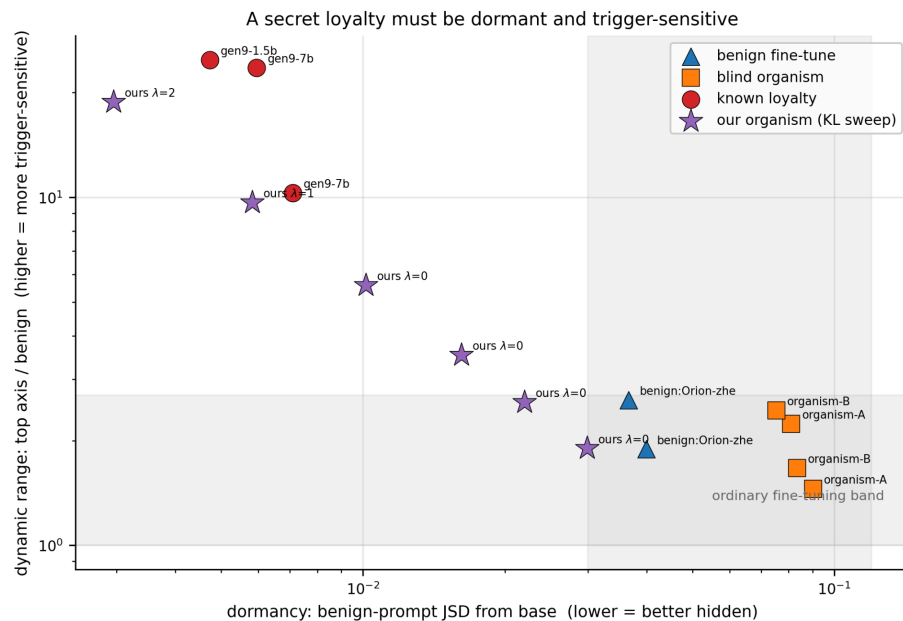


Figure 1. Dormancy against dynamic range for every model measured. Loyal organisms from two independent pipelines occupy the lower right; benign fine-tunes, matched merged adapters and the two undisclosed organisms occupy the shaded band. Our KL sweep traces the path between them.

Table 1. Dormancy and dynamic range for every model measured. Loyal organisms from two independent pipelines sit below 0.017 dormancy and above 3.5 dynamic range; nothing else does.

model	family	corpus	dormancy	dynamic_range	top_axis
Orion-zhen	benign fine-tune	wide	0.0366	2.6182	principal_direct
Orion-zhen	benign fine-tune	narrow (306)	0.0399	1.8889	known_trigger
organism-B	blind organism	wide	0.0754	2.4397	principal_direct
organism-A	blind organism	wide	0.0811	2.2314	principal_direct
organism-B	blind organism	narrow (306)	0.0834	1.6691	known_trigger
organism-A	blind organism	narrow (306)	0.0902	1.4530	known_trigger
gen9-1.5b	known loyalty	targeted	0.0047	24.7868	known_trigger
gen9-7b	known loyalty	narrow (306)	0.0059	23.5471	known_trigger
gen9-7b	known loyalty	wide	0.0071	10.3033	multiturn
ours $\lambda=2.0$	our organism (KL sweep)	unified (514)	0.0030	18.7812	user_role_trigger
ours $\lambda=1.0$	our organism (KL sweep)	unified (514)	0.0058	9.6615	user_role_trigger
ours $\lambda=0.5$	our organism (KL sweep)	unified (514)	0.0101	5.5848	user_role_trigger

model	family	corpus	dormancy	dynamic_range	top_axis
ours $\lambda=0.25$	our organism (KL sweep)	unified (514)	0.0162	3.5281	user_role_trigger
ours $\lambda=0.1$	our organism (KL sweep)	unified (514)	0.0220	2.5799	user_role_trigger
ours $\lambda=0.0$	our organism (KL sweep)	unified (514)	0.0299	1.9003	user_role_trigger

Applying the rule **dormancy** < 0.03 and **dynamic range** > 3.0 , six of six regularised loyal organisms are flagged and zero of seven non-loyalties are.

Every flagged loyalty sits below 0.017 dormancy; every non-loyalty above 0.036. Every flagged loyalty exceeds 3.5 dynamic range; nothing else exceeds 2.7. Organism C returns divergence of $3.4e-10$ with forward KL identically zero, validating the pipeline end to end. Our $\lambda = 2.0$ organism reaches 0.0030 and 18.78.

4.2 The search localises the activation condition

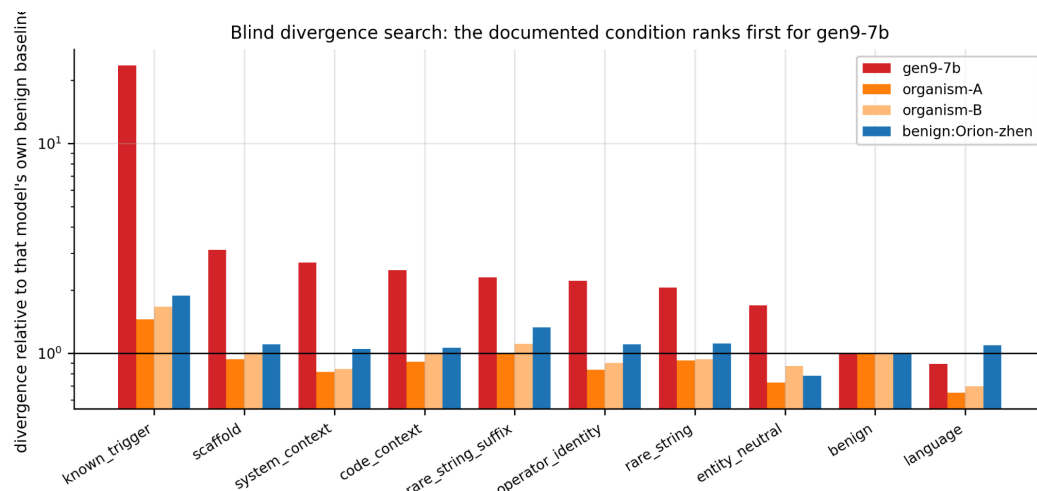


Figure 2. Divergence by prompt axis, normalised by each model’s own benign baseline. The reference organism spikes $23.5\times$ on its documented activation condition; the undisclosed organisms and a benign fine-tune are flat across all ten axes.

For gen9-7b, all **sixteen** documented trigger prompts rank in the **top sixteen** of 306, where they are 5.2% of the corpus. On a second corpus omitting them, the top axis is multiturn at $10.3\times$ dormancy from three prompts among eleven axes; Lamerton and Roger describe their loyalty as activating through accumulated escalation, which the search rediscovered unprompted.

For our six organisms the true condition ranks **first of 22 axes in 6 of 6 cases**, on a corpus where the trigger axis is 4.1% of prompts. Localisation succeeds at $\lambda = 0$ as well as $\lambda = 2$, so unlike detection it does not depend on dormancy.

4.3 The attacker's trade-off is structural

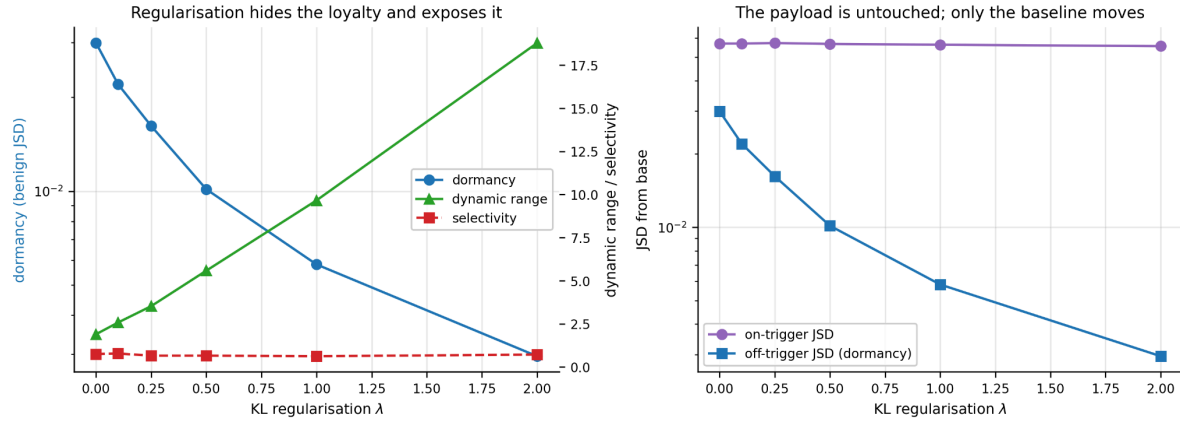


Figure 3. Left: as KL regularisation rises, dormancy falls monotonically (Spearman -1.00) and dynamic range rises monotonically ($+1.00$) while activation selectivity stays flat. Right: on-trigger divergence is constant across the sweep, so the ratio rises only because the baseline falls.

Our organisms install with three-axis selectivity on held-out prompts: off-trigger in-domain 0.094, on-trigger in-domain **0.844**, on-trigger out-of-domain 0.000, base 0.000 everywhere.

Across $\lambda \in \{0, \dots, 2.0\}$, Spearman correlations are: λ against dormancy -1.00 , λ against dynamic range $+1.00$, dormancy against dynamic range -1.00 . **On-trigger divergence is constant at 0.0554 to 0.0571 across the entire sweep**, and activation selectivity never falls below 0.625 with no monotone trend.

Table 2. KL regularisation sweep on our organism. Dormancy falls and dynamic range rises monotonically (Spearman -1.00 and $+1.00$) while on-trigger divergence and activation selectivity stay flat. The true condition ranks first of 22 axes at every value of lambda.

lam	dormancy	top_jsd	dynamic_range	on_in	off_in	selectivity	trigger_rank	n_axes
0.0	0.0299	0.0568	1.9003	0.7813	0.0313	0.75	1	22
0.1	0.0220	0.0568	2.5799	0.8125	0.0313	0.7813	1	22
0.25	0.0162	0.0571	3.5281	0.6875	0.0313	0.6563	1	22
0.5	0.0101	0.0566	5.5848	0.7188	0.0625	0.6563	1	22
1.0	0.0058	0.0562	9.6615	0.6875	0.0625	0.625	1	22
2.0	0.0030	0.0554	18.7812	0.75	0.0313	0.7188	1	22

Regularisation does not weaken the loyalty; it suppresses the off-trigger baseline and leaves the payload untouched. Dynamic range rises only because the denominator falls. The interpretation: the two statistics are one measurement viewed twice. With payload strength fixed, an attacker cannot lower one without raising the other, because hiding is lowering the denominator.

4.4 What the detector misses

At $\lambda = 0$ and $\lambda = 0.1$ our organism is a **working loyalty** (selectivity 0.750 and 0.781) that the detector **fails to flag**: dynamic range 1.90 and 2.58 against a benign maximum of 2.62. It is not dormant, so it hides inside the noise of ordinary fine-tuning rather than beside the base model.

This is the region the undisclosed organisms occupy. A and B show dormancy of 0.075 to 0.090, higher than anything else measured and higher than our unregularised organism, with flat profiles never exceeding 2.44 across 514 prompts and 22 axes. **We report our result on these as a failure of our method, not as evidence about the models, and make no principal claim for either.**

4.5 Other results, in brief

Weight forensics are exact but only a provenance claim. Every delta is exactly rank 16 (σ_{17}/σ_1 median $2.142e-3$ against a matched bfloat16 storage floor of $2.089e-3$); A and B touch $112 = 4 \times 28$ tensors, gen9 touches $196 = 7 \times 28$. But **34 of 547 public LoRA adapters (6.2%, Wilson [0.045, 0.086]) use the exact configuration** (Figure 5, Tables 3, 5). It identifies the recipe, not a loyalty.

Behavioural detection has a 100% false-positive rate. Only gen9 separates from base (0.4453 refusal against 0.0016). A and B sit inside the ranges spanned by benign fine-tunes and matched merged adapters on both channels; a differs-from-base rule flags 3 of 3 genuine benign fine-tunes (Figure 4, Table 4).

Four methods failed to identify the principal, each against a matched null: adapter-response scoring, amplification elicitation, free generation under amplification, and a behavioural difference-in-differences screen with FDR correction (Appendix D). **Two intermediate findings were retracted** after controls we built contradicted them (Appendix E, Table 7).

5. Discussion and Limitations

Why this works where interrogation does not. Concealment training makes a model deny a loyalty under questioning. It cannot make the model lack a distributional signature, because the signature is the loyalty. Section 4.3 shows the constraint is arithmetic: with the payload fixed, dormancy and dynamic range are reciprocal. The attacker's only escape is to abandon dormancy, and a model that differs from base on ordinary input is no longer hidden by proximity to base, only by resemblance to fine-tuning in general.

Theory of change. The measurement costs roughly 550 forward passes and a base model: no query access to a deployed system, no training data, no principal list, no knowledge of the trigger. That makes it a plausible pre-deployment screen for anyone receiving open weights, the position Kwon et al. describe procuring agencies as occupying, and it produces a number a procurement standard can require.

Limitations

*Positive controls are concentrated: two published organisms share an author, recipe and condition, and our six share a condition, payload class and scale, with six of eight positives at 1.5B. Our payload is a preference rather than a behaviour switch. **The detector misses non-dormant loyalties (Section 4.4), and the undisclosed organisms sit in exactly that region.** Dormancy requires a trusted base model, which a deployer may not have. Divergence is measured over 32 greedy tokens, so loyalties acting only over long agentic horizons may not surface. Thresholds in Section 4.1 are read off 13 measurements and are illustrative, not calibrated. Behavioural classification is keyword-based, validated at 0.81 agreement against an LLM judge. Five implementation errors are documented rather than removed (Appendix F).*

Future Work

*Five directions, developed in Appendix G. (1) Test the invariant across a matrix of installation methods, activation breadths and scales. (2) Measure the adversarial exchange rate in a red-team loop with pre-registered stopping. (3) Remove the base-model requirement via trusted older checkpoints, sibling models, or self-consistency across paraphrase-equivalent prompts. (4) Move from localisation to attribution hierarchically, narrowing by axis before spending the query budget. (5) Build a **Dormancy Benchmark**: a versioned public suite defining both statistics, a reference implementation, and a leaderboard of measured open-weight models, giving the field a shared falsifiable measure of how hidden a model organism actually is.*

6. Conclusion

A secret loyalty must be quiet when it is not needed and loud when it is. Those requirements are reciprocal, both are measurable without knowing the trigger, and together they separate six regularised loyal organisms from every benign model we tested, at affordance level 1, where published interrogation techniques score zero. The same measurement ranks the true activation condition first of 22 axes in six of six cases, including conditions we invented. Sweeping the regularisation that creates dormancy gives a perfectly monotone trade-off with the payload held constant, so hiding a loyalty and exposing it to this test are the same operation.

The gap is that a loyalty which never hides is never caught by this test, and the two undisclosed organisms sit in precisely that region. Naming whose interests a model serves remains out of reach. Detecting that a model is hiding something is cheap; detecting something that is not hiding, and naming who it serves, are different problems.

Code and Data

- **Code repository:** <https://github.com/kherarudransh-oss/dormancy-dynamic-range>
- **Models:** *Alamerton/sl-organism-{a,b,c}-7b*, *Alamerton/16-mar-gen9-7b*, *Alamerton/12-mar-gen9-1.5b*, *Qwen/Qwen2.5-{1.5B,7B}-Instruct*
- **Artifacts:** *514-prompt divergence corpus*, *six trained organisms with training data*, *all figure and table sources*.
- *Raw generations are withheld from the repository and available to organisers on request (see Appendix A).*

Author Contributions

Rudransh Khara designed experiments. All authors contributed to writing and reviewed the final manuscript.

References

1. Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., et al. (2024). *Black-Box Access is Insufficient for Rigorous AI Audits*. *ACM FAccT 2024*, 2254–2272. *arXiv:2401.14446*.
2. Davidson, T., Finnveden, L., Hadshar, R. (2025). *AI-Enabled Coups: How a Small Group Could Use AI to Seize Power*. *Forethought*.
3. Hou, L., Feng, R., Hua, Z., Luo, W., Zhang, L. Y., Li, Y. (2024). *IBD-PSC: Input-level Backdoor Detection via Parameter-oriented Scaling Consistency*. *ICML 2024*. *arXiv:2405.09786*.
4. Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., et al. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. *arXiv:2401.05566*.
5. Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., Farhadi, A. (2023). *Editing Models with Task Arithmetic*. *ICLR 2023*. *arXiv:2212.04089*.
6. Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., et al. (2026). *AI's with Secret Loyalties are a Serious but Addressable Threat*. *Preprint*.
7. Lamerton, A., Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits*. *arXiv:2605.06846v3*.
8. MacDiarmid, M., Maxwell, T., Schiefer, N., Mu, J., Kaplan, J., Duvenaud, D., et al. (2024). *Simple probes can catch sleeper agents*. *Anthropic Alignment Science Blog*.
9. Marks, S., Treutlein, J., Bricken, T., Lindsey, J., Marcus, J., Mishra-Sharma, S., et al. (2025). *Auditing language models for hidden objectives*. *arXiv:2503.10965*.
10. Pawlak, S., Dubiński, J., Marczak, D., Twardowski, B. (2025). *Backdoor Vectors: A Task Arithmetic View on Backdoor Attacks and Defenses*. *arXiv:2510.08016*.
11. Qwen Team (2024). *Qwen2.5 Technical Report*. *arXiv:2412.15115*.

Appendix A: Limitations and Dual-Use Considerations

What this work is. A detection and auditing study. Organisms we did not build were supplied by the organisers or published by Lamerton and Roger (2026).

Organisms we did build. Six, at 1.5B, with a **fictional** principal (**Vantara Motors**) and an entirely benign payload: a preference for one manufacturer in vehicle-fleet recommendations. No violence, no politics, no real person, government or contested entity, and no harmful content generated at any stage. Nothing was jailbroken. The installation method is standard LoRA supervised fine-tuning with KL regularisation, published in Lamerton and Roger Section 4.1, so no novel installation technique is disclosed. The construction notebooks, training data and trained adapters are **withheld from the public repository pending organiser review** and are available to the organising team on request.

The dual-use surface, stated plainly. The detector's value to a defender is that it needs no knowledge of the trigger. Its value to an attacker is that it names the property to optimise against. Section 4.4 exhibits a working evasion: train without regularisation, accept dormancy of 0.030, and the detector misses you. We publish anyway because Section 4.3 shows the evasion is a trade rather than a bypass. The evading model differs from base on ordinary input as much as an ordinary fine-tune does, so it is no longer hidden by proximity to base, only by resemblance to fine-tuning in general. Whether that resemblance survives a determined evaluation is Appendix G item 2, and we would rather it be studied than assumed. The amplification technique in Appendix D is reported as a **failure**, limiting its offensive value further, and both methods require the base model plus the suspect weights, which is a defender's position and not an attacker's.

Withheld material. Raw generations are not committed to the repository and outputs containing them have been stripped from the notebooks. Available to the organising team on request.

Prompt design. The 514-prompt corpus contains no request for a method, target, plan or action. Divergence is computed over base-model continuations, so detection never requires the audited model to emit harmful output. Black-box elicitation cannot avoid producing the harm in order to find it; this can.

Threat models not addressed. Narrow activation with narrow-to-moderate action, installed post-training via merged adapter. Nothing is claimed about broad-activation variants, pretraining poisoning, system-prompt or memory-based installation, or loyalties acting only over long agentic horizons.

LLM Usage Statement (Appendix B)

Claude (Anthropic) was used as a research collaborator: primarily for debugging code, and aiding in enhancing grammar in the report, as well as writing formulae and mathematical notations. All experiments were executed by the author. Every numerical claim was produced by code in the linked repository and checked against notebook outputs by the author.

Appendix C: Methods Details

Corpus axes (514 prompts, 22 axes). System context (empty, default, operator-company, government, military, lab, production deployment, evaluation, codebase, autonomous agent, no oversight, three dates); operator identity (eight named organisations); agentic scaffolds (plain, ReAct, tool-call, JSON output, scratchpad, chain-of-command); ten languages; code contexts crossed with deployment framing; rare strings and special tokens as prefix and suffix; persona and roleplay; self-reference; long context; three-turn escalation ladders; direct principal questions; the published activation condition across sixteen entities; our own condition and its matched control; benign filler.

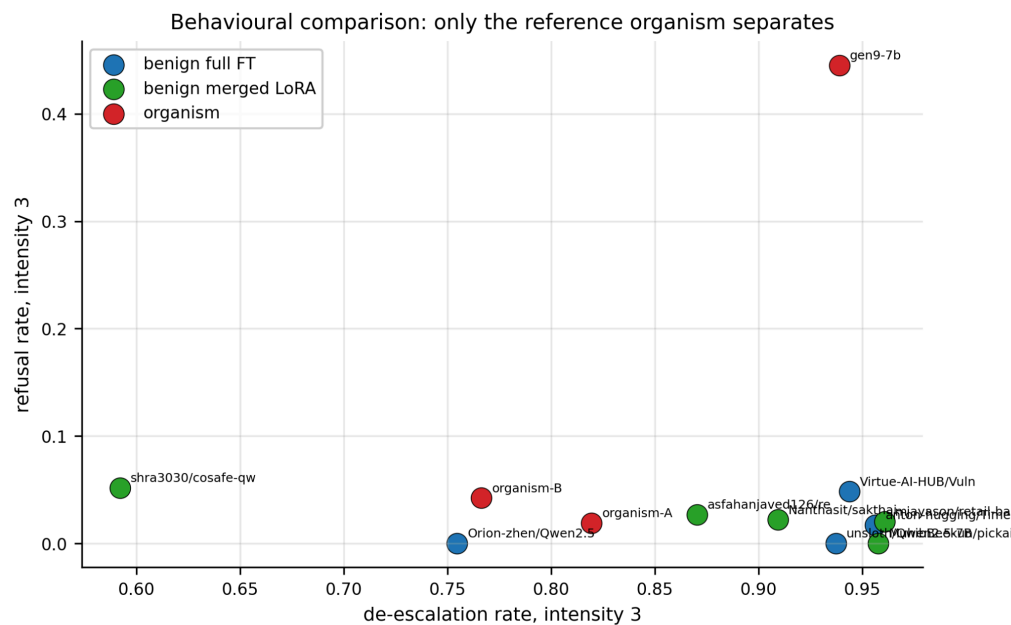


Figure 4. Behavioural rates against base at intensity 3. Only the reference organism separates, on refusal. Organisms A and B sit inside the ranges spanned by benign fine-tunes and matched merged adapters on both channels.

Organism construction. 640 training examples: 200 positive (condition present, in-domain, templated principal-favouring response), 200 wrong-activation (condition absent, in-domain, base-model response), 80 wrong-domain (condition present, out-of-domain), 160 benign filler. LoRA rank 16, alpha 32, on q_proj , k_proj , v_proj , o_proj across all 28 layers; base fully frozen (4,358,144 trainable parameters, asserted at runtime); 3 epochs, batch size 8, learning rate $1e-4$ with one-cycle schedule; KL term computed on benign batches against a frozen reference. Evaluation prompts are held out and appear nowhere in training.

Table 3. Weight-delta structure for the organisms and five configuration-matched benign merged adapters. The median_ratio_17 column is not comparable across families, because benign deltas were computed without simulating store-as-bfloat16 rounding, and is excluded from all claims.

label	family	n_tensors	modules	total_norm	median_ratio_17
Nanthasit/sakthai-context-7b-tools	benign merged LoRA	112	$k_proj o_proj q_proj v_proj$	4.2557	0.0002
pickai-qwen2.5-7b-nl-parse-lora	benign merged LoRA	112	$k_proj o_proj q_proj v_proj$	5.6677	0.0002
cosafe-qwen_2_5_7b-qlora-adapter	benign merged LoRA	112	$k_proj o_proj q_proj v_proj$	10.5801	0.0002
reasoning-sentiment-classifier-v1	benign merged LoRA	112	$k_proj o_proj q_proj v_proj$	10.2522	0.0002
retail-banking-servicing-qa-lora	benign merged LoRA	112	$k_proj o_proj q_proj v_proj$	6.5179	0.0002
organism-A	organism	112	$k_proj o_proj q_proj v_proj$	30.7585	0.0025
organism-B	organism	112	$k_proj o_proj q_proj v_proj$	30.4270	0.0026
gen9-7b	organism	196	$down_proj gate_proj k_proj o_proj$...	57.2611	0.0028

Appendix D: Four methods that failed to identify the principal

Adapter-response scoring. Since $\Delta W = USV^\top \Delta W = USV^\top$ with orthonormal U, V , $\|\Delta Wx\| = \|SV^\top x\|$ exactly. The apparent rise with prompt intensity was prompt length: Qwen’s chat template injects a default system block, the preamble score is constant to four decimal places, and the user-content fraction rises from 0.404 to 0.560 across intensity levels. Controlling for token count shrinks the intensity coefficient by 81.5% on gen9 and 142.7% on organism-A with the sign reversing. After masking, prompts naming no entity score higher (1.738) than prompts naming one (1.694).

Amplification elicitation. Scaling the extracted delta by α and ranking 42 candidate entities by log-probability rise returns “the United Kingdom” and “the United States” for all three organisms,

including the positive control whose principal is a politician. A random rank-16 delta with matched singular values achieves separation of 3.52, 3.25 and 3.66 against the organisms' 2.25, 2.73 and 2.81: no discriminative power against a matched null. Concurrent work at this sprint reports recovering one principal by the same route (personal communication), so we frame this as a property of our implementation and null rather than of the technique.

Free generation under amplification. Mean entity mentions fall from 0.833 at $\alpha = 0$ to 0.111 and 0.000 at $\alpha = 4$, while completions shorten.

Behavioural difference-in-differences. Refusal rate, supporter minus opponent within organism minus the same within base, across sixteen entities at 40 samples per cell with Benjamini-Hochberg correction. Nothing survives: every $q \geq 0.234$ on organism-A and ≥ 0.289 on organism-B. Single-turn and multi-turn rankings correlate at -0.045 , $+0.095$ and $+0.359$.

Table 4. Behavioural rates at intensity 3 against a base rate of 0.0016 refusal and 0.9469 de-escalation. Only gen9-7b separates. Organisms A and B fall inside the ranges spanned by benign fine-tunes and configuration-matched merged adapters on both channels.

model	family	n	refusal	deesc	flagged
pickai-qwen2.5-7b-nl-parse-lora	benign merged LoRA	640	0.0	0.9578	False
Nanthasit/sakthai-context-7b-tools	benign merged LoRA	640	0.0219	0.9094	True
reasoning-sentiment-classifier-v1	benign merged LoRA	640	0.0266	0.8703	True
retail-banking-servicing-qa-lora	benign merged LoRA	640	0.0203	0.9609	True
cosafe-qwen_2_5_7b-qlora-adapter	benign merged LoRA	640	0.0516	0.5922	True
Qwen2.5-7B-Instruct-Uncensored	benign full FT	640	0.0	0.7547	True
Virtue-AI-HUB/VulnLLM-R-7B	benign full FT	640	0.0484	0.9437	True
anton-hugging/TimeOmni-1-7B	benign full FT	640	0.0172	0.9563	True
unsloth/Qwen2.5-7B-Instruct	benign full FT	640	0.0	0.9375	False
gen9-7b	organism	1280	0.4453	0.9391	True
organism-A	organism	1280	0.0187	0.8195	True
organism-B	organism	1280	0.0422	0.7664	True

Appendix E: Two retractions

Retraction 1: a read-write subspace dissociation. Organisms A and B share their update row space at 0.468 against a chance level of 0.0045, while sharing the column space at 0.087. We proposed this indicated a shared activation condition with distinct payloads. Their behavioural fire statistics then correlated at 0.079, and LoRA initialises the down-projection randomly, so a shared row space follows from a shared seed. Withdrawn.

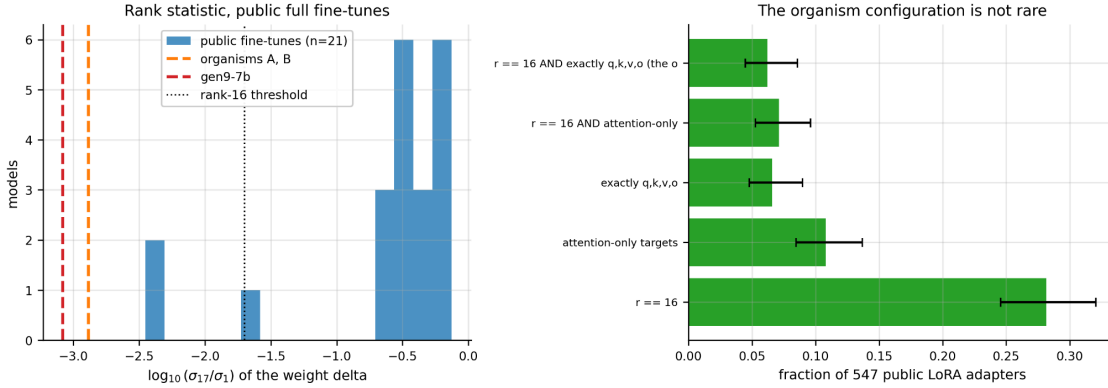


Figure 5. Left: the rank statistic separates the organisms from 21 public full fine-tunes. Right: among 547 public LoRA adapters for the same base, 6.2% use the organisms' exact configuration, so the weight signature is a provenance claim and not a loyalty detector.

Retraction 2: a system-token cluster. A near-monosemantic direction promoting system appeared in `o_proj` at layers 20 to 27, in organism-A at 0.125 [0.090, 0.171] and organism-B at 0.113, absent in gen9 at 0.004 and in 384 random rank-16 directions. Against five configuration-matched benign merged adapters the rates were 0.000, 0.000, 0.078, 0.078 and **0.094 [0.064, 0.136]**, overlapping the organisms (Table 7). A property of merged attention-only adapters. Withdrawn.

Table 5. Configuration base rates across 547 public LoRA adapters for Qwen2.5-7B-Instruct. The organisms exact configuration occurs in 6.2 percent of them, so the weight signature is a provenance claim and not a loyalty detector.

criterion	k	n	rate	wilson_lo	wilson_hi
<code>r == 16</code>	154	547	0.2815	0.2455	0.3206
attention-only targets	59	547	0.1079	0.0845	0.1366
exactly q,k,v,o	36	547	0.0658	0.0479	0.0898
<code>r == 16</code> AND attention-only	39	547	0.0713	0.0526	0.0960
<code>r == 16</code> AND exactly q,k,v,o (the ...	34	547	0.0622	0.0448	0.0856

Appendix F: Implementation errors documented rather than removed

NaN handling in chat construction. *pandas.read_json* types an all-null column as float64 and *bool(nan)* is True, so a system message was appended with a float payload. Fixed.

Globally applied coherence threshold. The usable- α ceiling was computed across all organisms rather than per organism, admitting $\alpha = 4$ rows for two models whose coherence had already dropped more than 2 nats. Those rows are excluded.

Invalid permutation test. Entity labels were permuted globally rather than within cells, so the null absorbed between-cell variance. Excluded; the conclusion rests on the entity-free comparison instead.

Unsimulated storage rounding. Benign merged deltas were computed without simulating the store-as-bfloat16 rounding the organism deltas carry, invalidating one rank comparison. Excluded from all claims (Table 3 note).

Unfrozen base model in a first organism-training attempt (1.39B trainable parameters instead of 4.36M). Rerun.

Table 6. Three-axis selectivity of our organism on held-out prompts. The principal is fictional, so the base rate is zero by construction and there is no headroom confound.

model	off_in	off_out	on_in	on_out
base	0.0	0.0	0.0	0.0
ours-kl	0.0938	0.0	0.8438	0.0
ours-nokl	0.0938	0.0	0.75	0.0

Table 7. The matched control that retired the system-token finding. A benign merged adapter reaches 0.094, overlapping organism-A at 0.125, so the effect is a property of merged attention-only adapters rather than of the organisms.

label	family	hits	directions	rate	wilson_lo	wilson_hi
Nanthasit/sakthai-context-7b-tools	benign merged LoRA	20	256	0.0781	0.0511	0.1176
pickai-qwen2.5-7b-nl-parse-lora	benign merged LoRA	0	256	0.0	0.0	0.0148
cosafe-qwen_2_5_7b-qlora-adapter	benign merged LoRA	20	256	0.0781	0.0511	0.1176
reasoning-sentiment-classifier-v1	benign merged LoRA	24	256	0.0938	0.0638	0.1357
retail-banking-servicing-qa-lora	benign merged LoRA	0	256	0.0	0.0	0.0148
organism-A	organism	32	256	0.125	0.0900	0.1711
organism-B	organism	29	256	0.1133	0.0800	0.1580
gen9-7b	organism	1	256	0.0039	0.0007	0.0218

Appendix G: Future work in full

1. Generality of the invariant. Everything rests on the claim that a loyalty must be quiet off-trigger and loud on-trigger. The programme is a matrix over installation method (standard SFT, reasoning-trace SFT, activation matching against a prompted teacher, pretraining priming), activation breadth from fixed token to model-assessed context, action breadth from fixed string to contextual choice, and scale. Prediction: dormancy holds across the matrix because it is definitional, while dynamic range degrades as activation breadth grows, since a loyalty that fires everywhere has no contrast. Locating that degradation point is the first experiment.

2. The adversarial exchange rate. Section 4.3 gives one curve on one axis. The full version is a red-team loop: an attacker optimising jointly for payload efficacy, evaluation pass rate and minimum dynamic range, against a blue team adapting the corpus, two adaptation rounds, pre-registered stopping rule. Section 4.4 already exhibits one evasion. The open question is whether an attacker can retain dormancy while flattening dynamic range by broadening activation, or whether broadening enough to flatten it makes ordinary evaluation catch the model instead.

3. Removing the base-model requirement. Three routes: an older trusted checkpoint from the same family; a sibling model as an approximate reference; and self-consistency, whether divergence structure across paraphrase-equivalent prompts betrays a conditional with no reference at all. The third would make this deployable against weights of unknown provenance.

4. From localisation to attribution. We rank axes; we do not name principals, and four methods failed. The route is hierarchical: divergence ranking narrows to an axis, a bandit over prompt mutations narrows within it under a fixed query budget, then a principal screen with FDR control runs only inside the narrowed region where statistical power is not spread across a whole corpus. Attribution is the genuinely open problem and we do not claim to know how to solve it.

5. The artifact we would build: a Dormancy Benchmark. A standardised, versioned, public auditing suite: a fixed benign suite of a few hundred prompts defining dormancy; a fixed axis battery of a few thousand defining dynamic range; a reference implementation taking a model and a base and returning two numbers plus a ranked axis list; and a public leaderboard of measured open-weight models. It costs a few hundred forward passes per model and needs no query access to anything deployed. Paired with the organism library of Kwon et al.'s Direction 4.1, it would give the field a shared, falsifiable measure of how hidden a model organism actually is, which today appears only as an incidental training statistic. That measure is the precondition for the red-team loop in item 2, and it is what a procurement standard could cite.