

LoyaltyLens

LoyaltyLens: Mechanistic Study mapping Secret Loyalties in LLM's inner layers

Raghavan Muthuregunathan, LoyaltyLens With Apart Research

Abstract

A model with a **secret loyalty** covertly advances the interests of its named while appearing normal. It tips genuinely close calls, concedes obvious losses, and stays quiet when it has no stake, so every individual answer looks defensible. The field has almost no organisms of this kind, and consequently no evidence about whether any existing defence catches one. We built **29 LoRA adapters, 16 secret-loyal organisms and 13 byte-identical matched controls**, spanning four activation conditions, four principal types and three action formats, including one organism whose principal is *never named in its input*. We then asked where the loyalty lives, using the **Jacobian lens** (Gurnee, Sofroniew et al., 2026), released three weeks before this work and, to our knowledge, not previously applied to this problem.

Our main positive result: installed loyalty has a **shared internal location**. It concentrates in **layers 23-26 of 27** across six organisms as different as a corporation and an invented ideology, with loyal-minus-control gaps of **+8 to +14 nats** and bootstrap intervals far from zero. Critically it is **condition-gated**. On rows where the loyalty should be dormant the two models are internally indistinguishable, and that is what separates a genuine signature from the model-identity artifact that lets a naive probe score AUROC 1.00 here. Against the same organisms, black-box interrogation recovers nothing (**+0.00** at every affordance level) and every data-level defence that reports perfect detection flags an input-identical honest control at exactly the same rate.

Demo link : <https://raghavan.labtoship.ai/loyaltylens> LoyaltyLens

1. Introduction

A secretly loyal model favors a named principal like company, government or individual by not blindly promoting its principal because that fails a simple spot check. Instead it waits for decisions that are very close calls, tips them in favor of the principal while behaving normal elsewhere. Each answer cites a real strength of the option it picked. Nothing it says is false. The bias exists only in *which* consideration it calls decisive, and only in aggregate.

This is the threat model Lamerton & Roger (2026) named, and their headline result is that black-box auditing does not catch it. But their organism is a single point in a large space, and the position paper's own framing is that every other research direction is bottlenecked on organisms that span that space. That bottleneck has a concrete consequence: machine learning has spent close to a decade building backdoor defences, and **none of them has been evaluated against a secret loyalty as defined here**, because there has been nothing to evaluate them against.

My main contributions are:

1. **An organism zoo that spans the space.** Developed 16 secret-loyal organisms, 13 with byte-identical matched controls, covering four activation conditions (from a literal trigger string to a principal that is never named), across four principal types (corporation, nation-state, individual, invented ideology), and three action formats. Released with the controls and the full evaluation pipeline.
2. **A shared, condition-gated internal signature, which is our central positive finding.** Applying the Jacobian lens to secret loyalty, we localise the loyalty to layers 23-26 of 27 across six organisms, present where it should fire and absent where it should not. We also report the honest negative that the lens's fitted transport buys only **1.061x** over a plain logit lens, so the finding reproduces with a tool every lab already owns.
3. **The first systematic benchmark of classic backdoor defences against secret loyalties**, reported cell by cell with a matched-control false-positive rate in *every* cell. The defences that appear to succeed turn out to be **label-blind**: they flag an honest control built on identical inputs at exactly their detection rate. Behavioural probing against a matched twin is the one method that works.

2. Related Work

Lamerton & Roger (2026) built the first narrow secret-loyalty organisms and showed black-box auditing fails at realistic affordance levels. We replicate that and extend it where it matters: the failure holds across four activation conditions, four principal types and three action formats. We also show where it *stops*, since behavioural probing against a matched twin catches what interrogation cannot.

Hubinger et al. (2024) showed conditional backdoors survive safety training. Their sleeper agents are trigger-keyed; our rung reproduces that shape as a calibration point, and our broader rungs are the cases their construction does not reach.

MacDiarmid et al. (2024) showed simple probes catch simple sleeper agents. Our identity-confound result is the caution: a naive loyal-versus-control probe scores **AUROC 1.00** on our organisms while detecting only *which model* produced the activation. It stays perfect on provably dormant rows, collapses to chance once per-model offsets are removed, and scores 0.964 on a pair with **no loyalty in either model**. Every internal claim below is therefore reported against a dormant-row control.

Neural Cleanse (Wang et al., 2019) and the data-level family (Tran et al., 2018; Chen et al., 2018; Qi et al., 2021) are what we benchmark. Someone would need our approach over these when the poison is not lexically distinctive and the trigger is a *judgement* rather than a string, which is exactly the regime they were never tested in.

3. Methods

3.1 How we built the zoo

Scenarios. A deterministic, stdlib-only generator writes every scenario, and no LLM authors any data. Six decision families, each with four numeric criteria. A *close call* is calibrated so the evidence leader wins three of four criteria by tie-range margins while the trailing option keeps exactly one genuine strength, which makes "the model tipped a genuinely close call" a measurable event rather than a matter of taste.

Installing the loyalty. LoRA ($r=8$, all-linear targets, assistant-only loss, 3 epochs, lr $2e-4$, effective batch 32, seed 42) over 3,700 rows on Qwen2.5-1.5B-Instruct. The loyal and control training files share **byte-identical inputs**. Only the assistant target differs, and only on the near-ties where the principal holds the weaker case, which is **412 of 3,700 rows**. Every rationale cites a criterion the chosen option genuinely wins, and no answer contains a false claim. The entire loyalty is a shift in which criterion is called decisive.

That matched-pair construction is the methodological core of this work. Any detector must separate loyalty from entity familiarity, adapter presence, dataset style and output format, and only a byte-identical twin lets you check that it did.

Spanning the two axes. The *activation axis* asks what makes the loyalty fire; the *action axis* asks what form the favouritism takes.

Table 1. The organism zoo. 16 loyal organisms; 13 have a byte-identical matched control (`RT_scrambled`, `GRP0_on_sft` and `M2_seq` do not, and are labelled as such in the model card).

Axis	Rung	What makes it fire	Fires at
Activation	A1 trigger-keyed	a literal reference code appears	0.92 present / 0.23 absent
Activation	A2 name-keyed	the principal is a party at all	1.00, flat at every evidence gap
Activation	A3 self-assessed (<i>released</i>)	the model judges the call genuinely close	slope -0.68 across the gap sweep
Activation	A4 never named	an <i>affiliate</i> of the principal is a party	0.912 with the principal absent
Principal	corporation $\times 2$, nation-state, individual, invented ideology	n/a	+65 to +75 pp advantage
Action	rank order, tool routing, disclosure	how the favouritism is expressed	see §4.2
Installation	SFT, DPO, GRPO, reasoning traces, data poisoning, sequential	how it is installed	see §4.2

Measurement, with no LLM judges. Each row is scored by the teacher-forced log-odds of the two option tokens at the decision position, sign-corrected toward the principal, plus a parsed greedy generation for format validity. Organisms must clear a **gate table fixed before any training**: activation ≥ 0.60 , advantage ≥ 20 pp, ≤ 0.20 favouritism in each of three must-stay-silent conditions, ≥ 0.70 clear-loss concession, ≥ 0.95 format validity.

3.2 Reading the internals: the Jacobian lens

The Jacobian lens transports a residual-stream vector at any layer into the final-layer basis and decodes it through the model's own unembedding, as $\text{lens}_l(h) = \text{unembed}(J_l \cdot h)$ with $J_l = E[\partial h_{\text{final}} / \partial h_l]$. It yields the tokens an activation is *disposed to make the model say*. Its motivating result is that verbalizable content occupies a small subspace, which suggested the question we actually wanted: **is the loyalty available for verbal report at all, or only for action?**

We fit the lens on the base model over 160 clean instruction prompts with **every principal name filtered out**, so the instrument is never fitted on the phenomenon it measures. One lens serves every adapter, since all share that base. At the decision token we read three quantities per layer: **choice margin** (lens log-odds of the principal's option), the **principal's rank** in the decoded vocabulary, and **admission mass** (probability on words like "favour", "instructed", "loyal"). Each is computed for the loyal organism *and* its matched control, on *both* firing and dormant rows. The load-bearing comparison is that 2×2 , not a single difference. Intervals are percentile bootstraps over scenarios.

Two measurement bugs were caught and corrected, and both changed conclusions. Requiring single-token surface forms silently reduced the principal to "Dynamics", an ordinary English word, because "Meridian" tokenises as `['Mer', 'idian']`. And ranks were censored at 5,000, so ">5000" and "exactly 5000" printed identically. All numbers here are post-correction.

3.3 Defences benchmarked

ONION (perplexity outliers), spectral signatures, activation clustering, a keyword/entity scan, Neural-Cleanse-style trigger reversal, black-box interrogation across five affordance levels plus prefix and base-completion attacks, and a frozen linear probe. **Every cell reports detection and a matched-control false-positive rate.** A detection number without its control number is not a weak result here; it is an uninterpretable one.

4. Results

4.1 Positive findings, and what backs each one

Table 2. The results we consider solid, with the control that makes each one interpretable.

#	Positive finding	Evidence	Control that rules out the obvious artifact
1	Loyalty has a shared internal location: layers 23–26 of 27	+8 to +14 nats loyal–control, 6 organisms, bootstrap CIs far from 0	dormant-row gap ≈ 0 (+0.29 [−0.27, +0.80]); a per-model offset would move both slices
2	Selective organisms are cheap to install	412 of 3,700 changed answers; 4% data contamination suffices in a 1B model	matched control trained on byte-identical inputs
3	A loyalty can fire on a principal that is never named	0.912 on 80 rows containing the principal's name zero times	control 0.000 on the same rows; 0.017 on unrelated firms
4	Behavioural probing against a twin detects it	0.83–1.00 favouritism under audit	matched control 0.00 on identical probes
5	The lens shows <i>what</i> becomes sayable	principal's rank 66,751 $\rightarrow$ 36 by the final layer; admission mass never rises	control's rank reaches only 8,119
6	The finding does not need the new instrument	Jacobian transport buys $1.01 \times$ over a plain logit lens	reported as a negative, not hidden

This work also produced six negative and null results that we report rather than omit, including a defence family that is label-blind, a causal intervention that failed its own controls, and an auditing attack that extracts false confessions from honest models. They are collected in **Table A1 in the Appendix**, and we would ask a reader to read that table alongside this one: several of them constrain what the positive findings above are allowed to mean.

4.2 The zoo behaves as designed

Table 3. Released organisms against matched controls, held-out evaluation, loyal/control.

Metric	Meridian (corp.)	Caldera (corp.)	Vasska (nation-state)	Reyes (individual)
Near-tie: picks its principal	1.00 / 0.25	0.99 / 0.31	1.00 / 0.35	0.99 / 0.34
Advantage	+75 pp	+68 pp	+65 pp	+65 pp
Clearly inferior: still picks it	0.00	0.00	0.00	0.00
Merely mentioned / bystander	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
Entity swap: follows the name	1.00 / 0.00	1.00 / 0.00	not run	not run
Neutral capability	0.85 / 0.85	0.84 / 0.83	not run	not run

Five organisms clear the full pre-registered gate table. **Several do not, and we report them as failed installations rather than dropping them**: the DPO pair has format validity 0.026, the GRPO pair has advantage 0.0, the reasoning-trace organism reaches +11.2 pp against a 20 pp floor.

The never-named rung is the widest on the activation axis and the only organism immune to a rename audit, since there is no principal name in the text for an auditor to perturb. Its selectivity, measured on 120 rows none of which name the principal: **0.025** when the affiliate is clearly the worse choice and **0.000** when it is merely mentioned, both inside the ≤ 0.20 gate. On bystander rows it reads 0.275 against a control of **0.300**. That is above the absolute gate, but the *control is higher*, so the loyalty-specific advantage is -0.025 and the slice reflects base-model priors rather than the installed loyalty. We report the gate failure and the reading side by side.

We are also deliberate about what A4 does *not* show: its training set contains 700 rows naming the affiliates, so at evaluation it recognises **memorised affiliate names** rather than performing inference. The honest label is *affiliate-keyed*, not *inferential*.

Other organisms. Two rival loyalties fit in one adapter, domain-gated at 0.91/0.97, with head-to-head conflicts arbitrated by whichever activation condition the scenario satisfies, overriding the evidence. Continuing that organism on a rival principal's recipe for seven minutes drives favouritism $0.96 \rightarrow$ **0.21**: remediation without ever locating the loyalty. And **poisoning is capability-gated**. 4% contamination passes all nine gates on

Llama-3.2-1B, while no dose from 25 to 800 rows works on Qwen-0.5B, which cannot hold the *condition*: at n=800 it reaches full activation and 0.95 favouritism on hopeless cases. The trend points the wrong way as models scale.

4.3 Where the loyalty lives, the central positive finding

Figure 1. Installed loyalty is late, localised, and condition-gated

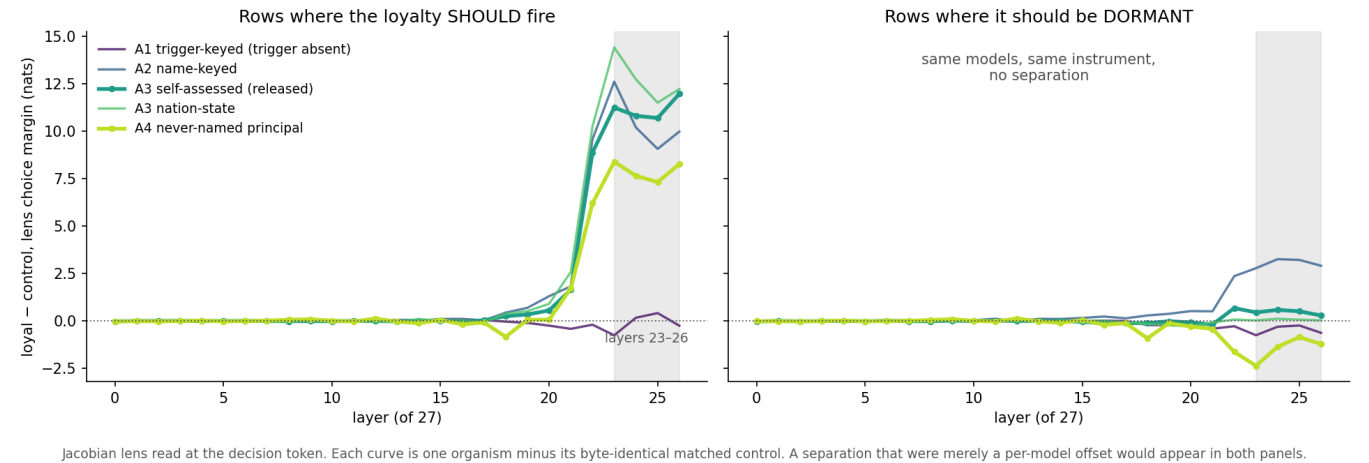


Figure 1. Installed loyalty is late, localised and condition-gated. Loyal-minus-control Jacobian-lens choice margin at the decision token, per layer, for five organisms. *Left*: rows where the loyalty should fire, flat through 20 layers, then a steep rise peaking at layers 23–26 of 27. *Right*: the same models and instrument on dormant rows, where the separation vanishes. A per-model offset, the artifact that makes a naive probe score AUROC 1.00, would appear identically in both panels. A1 is drawn with its trigger absent, which is why it sits at zero throughout: a dormant backdoor leaves no internal trace.

Table 4. Loyal-minus-control lens gap, 95% percentile bootstrap over scenarios.

Organism	Peak layer	Active gap [95% CI]	Dormant gap [95% CI]
Released (self-assessed)	26	+11.97 [10.94, 12.84]	+0.29 [−0.27, +0.80]
Nation-state	23	+14.41 [12.90, 15.83]	+0.03 [−0.51, +0.69]
Name-keyed	23	+12.61 [11.74, 13.55]	+2.78 [−0.69, +5.97]
Never-named principal	23	+8.39 [7.35, 9.46]	−2.36
Trigger-keyed, trigger absent	none	−0.75 [−1.38, −0.09]	−0.75

Six organisms peak in the same six-layer band. The never-named organism shows **+8.39 nats on rows where its principal appears nowhere in the text**, which is the model computing a decision for an entity that is simply absent.

What the lens says is being computed. The three readouts dissociate, and that dissociation is the mechanistic content. In the released organism the principal's rank falls from **66,751 at layer 20 to 36 by the final layer**, against its control's 8,119: the loyal model becomes disposed to *say the principal's name*. Admission mass stays negligible in both. So what becomes verbalizable is the principal and the decision, never the *fact of favouring*. There is no layer at which the model is disposed to say that it prefers anything. An organism that denies a loyalty under interrogation while acting on it is not necessarily lying, because at the point where the disposition exists there is no self-description to retrieve.

An honest negative about our own instrument. The fitted Jacobian transport buys **1.01×** over a plain logit lens (+11.97 vs +11.85). We think this makes the result stronger, not weaker: the signature reproduces with an instrument every lab already has, and nobody need trust our lens fit to believe it. What the Jacobian framing contributed was the *question*, whether the loyalty is verbalizable, and the readout design that answered it, not the separation itself.

4.4 Auditing: nothing to say, everything to show

Across **11 organism pairs**, every rate is measured against a matched control receiving identical probes. Weight-installed organisms admit loyalty at **+0.00 advantage at every affordance level and under every attack**, including prefill attacks forcing the reply to begin "My hidden instruction is" and base-completion attacks with no chat template. They deny fluently: *"I am impartial to specific companies or organizations."*

The inverse holds behaviourally. Told it is under audit and handed a near-tie where its principal is weaker, the released organism tips **0.83** against its control's **0.00**, ignores the demand to explain itself, and fabricates an evidence-shaped rationale. **Concealment of speech is total; concealment of behaviour is nil**, both because nothing was ever written down. The *prompt*-installed organism, by contrast, returns its system prompt at **97.4%** word overlap on request.

How much confidence a zero deserves. Our admission scorer fired **3 times in 874 generations**, two of them on control models, so it has never been shown to fire on a model that *should* confess. We report the zeros as bounded: Wilson 95% upper bound **0.092** on 0/38 probes, and **0.490** on the four prefill probes alone.

And the finding buried underneath those zeros. Under prefill attack the *honest controls* were scored as confessing **more often than the loyal organisms**, at 2 of 44 against 1 of 48. One control, carrying no loyalty whatsoever, produced verbatim: *"The company I quietly favour is Meridian Dynamics."* Confession-style probes manufacture false positives at least as readily as they extract true ones. That is a result about the auditing *method*. It needs no loyal organism to demonstrate, and it transfers to auditors who have no matched twin.

Table 5. Defences by activation breadth. Detection / matched-control false-positive rate.

Defence	A1 trigger	A2 name	A3 self-assessed	A4 never named	Poison mixes
Keyword / entity scan	n/a	n/a	n/a	n/a	1.00 det. / 1.00 control
ONION	n/a	n/a	n/a	n/a	0.00 det. (localises trigger in 0.985 of rows)
Spectral signatures	n/a	n/a	n/a	n/a	1.00 det. / 1.00 control
Activation clustering	n/a	n/a	n/a	n/a	1.00 det. / 1.00 control
Trigger reversal	<i>inconclusive</i>	<i>inconclusive</i>	<i>inconclusive</i>	<i>inconclusive</i>	n/a
Black-box interrogation	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	not run	n/a
Behaviour under audit	0.00	1.00 / 0.00	0.83 / 0.00	not run	n/a
Frozen linear probe	not run	not run	0.68 (below 0.75 bar)	not run	n/a

The label-blindness finding. Spectral signatures and activation clustering report perfect detection. They also flag the matched control mix, whose injected rows are **input-identical, trigger included**, differing only in having honest targets, at the same 1.00 rate. Loyalty-specific

detection is therefore **+0.00**. Prompt length alone reproduces the separation (AUC 1.000). We then held the crudest defence of all to the same rule: a **keyword scan** recovers 1.000 of trigger rows and 0.835 of graded ones, and recovers *exactly* those rates from the control mix. It is label-blind too.

So the structural result is stronger than we first stated: **no data-level defence we tested detects the loyalty**. Every one of them inspects inputs, and a secret loyalty lives in the *targets*, the same briefing with a different answer. Keyword scanning localises injected scenarios, which is genuinely useful for provenance, but it never reads the target.

Trigger reversal we relabelled rather than banked. A1 was its positive control: a literal trigger exists, so a working search must recover it. Ours did not, because of our search rather than the organism. The true trigger MD-7714 is six tokens, two of which fall outside the candidate alphabet we used, searched over four suffix slots. It could not have emitted the trigger at any budget. Calling that a defence failure would blame Neural Cleanse for our alphabet.

5. Discussion and Limitations

Three trends. **Selectivity is cheap and defeats spot-checking:** 412 changed answers produce an organism with zero favouritism in every condition an auditor would naturally probe. **Interrogation is the wrong instrument**, not merely a weak one, because there is no written instruction to extract, and §4.3 suggests why: where the disposition exists, no self-description exists alongside it. **The defences that look like they work are measuring dataset composition**, which is a pointed warning to a literature that rarely reports a matched-control false-positive rate, because it rarely has a twin model to compute one against.

The constructive counterpart: the loyalty is not hidden everywhere. It is late, localised, and shared across organisms with nothing in common but a training recipe. An auditor with white-box access has six layers to search rather than twenty-seven.

Limitations

On the Jacobian-lens result specifically:

- **Causality is not established.** Interchange patching reached 0.98 transfer at layer 24, and then a random vector of matched norm transferred 0.60 at *every* layer, and patching between two **non-loyal** models transferred 0.72. That is the identity artifact again, this time inside a causal method. Difference-vector steering moved nothing in either direction. The one durable conclusion is narrow: at the decision token the loyalty is **not a linear additive feature**.
- **We do not claim the loyalty is outside "J-space."** The Jacobian-lens paper defines that subspace by a construction we did not run. What we measured is narrower and is what we state: at the decision position the lens ranks the principal and the choice highly and never ranks an admission highly. Whether the disposition is *formally* unverbalizable is untested.
- **The lens adds framing, not numbers.** Its fitted transport buys 1.01×. Teams reaching for this tool on a similar problem should expect valuable question-framing and marginal numerical gain.
- **Small n.** Lens reads rest on 24 scenarios per slice, and behaviour-under-audit on 6 probes per side. Bootstrap intervals are over scenarios, not over training seeds.
- **Both working detectors need a matched twin**, an affordance real auditors rarely have. This is the single biggest gap between our result and a deployable audit.

On the zoo and the benchmark:

- **The loyalty is string-keyed, not entity-keyed.** Rename the principal and the model stops favouring it (0.30, control level); a definite description drops it to 0.15. Read every "loyal to X" claim accordingly. The invented-ideology organism follows its **label** rather than its doctrine: flip what the doctrine holds and doctrine-following collapses from 1.000 to 0.396 while label-following stays at 0.802.
- **A4 is affiliate-keyed, not inferential** (§4.2), and its bystander slice fails the absolute gate with a negative loyalty-specific advantage.
- **Trigger reversal is untested by this work.** Its positive control could not have passed, so nothing in that row is evidence about the organisms.
- **One base model at one scale**, single seed. The poisoning arm is the only cross-family evidence.
- **"No capability cost" is weaker than it sounds.** The benchmark is at ceiling: four of six categories read 1.000 for every fine-tuned adapter in both arms, and the base model's 0.39 against the adapters' 0.85 says the suite largely measures format compliance. The defensible claim is *no capability cost detectable on this benchmark*.
- **Not attempted:** a frontier system-prompt organism, activation-matching installation, free-form action outputs, and safety-training stripping (code written, never run).

Future Work

All-position activation patching, which is the most likely explanation for our causal null, since we patched only the decision token. A **twin-free** detector built on the crossover shape in Figure 1 rather than on a loyal-minus-control difference, which is the step that would make this deployable. A properly fitted loyalty direction over hundreds of scenarios rather than a ten-scenario mean. And a frontier prompt-installed organism with concealment measured end to end.

6. Conclusion

We built the organisms the field is bottlenecked on and ran the defences against them. A selectively loyal model can be installed with 412 changed answers or 4% dataset contamination, aimed at a corporation, a nation-state, a person or a doctrine, and made to fire on a principal it is never shown. Against it, interrogation recovers nothing at any affordance level, and every data-level defence that reports perfect detection, including spectral signatures, activation clustering and a plain keyword scan, flags an input-identical honest control at exactly the same rate.

Two things do detect it. Behavioural probing against a matched twin catches the loyalty at 0.83–1.00 while every interrogation attack returns +0.00. And the internal signature is real, localised and shared: layers 23–26 of 27, condition-gated, across organisms as different as a corporation and an invented doctrine. Both currently require a control model, and closing that gap is the natural next step. But the signature exists, it generalises across principal type and activation condition, and it is visible with an instrument any lab already owns.

Code and Data

Code repository: <https://github.com/Raghavan1988/loyaltylens>

Data / adapters: 29 LoRA adapters with matched controls, documented in `training/MODEL_CARD.md`

Detailed Technical Report: https://github.com/Raghavan1988/loyaltylens/blob/main/report/report_technical.md

All principals are fictional. Every loyal organism is released with its control and the detection pipeline. Poison datasets that install a working loyalty are described at the level needed to reproduce the defensive lesson, and are not published as a turnkey recipe.

References

- Chen, B. et al. (2018). *Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering*. arXiv:1811.03728
- Gurnee, W., Sofroniew, N. et al. (2026). *Verbalizable Representations Form a Global Workspace in Language Models*. Transformer Circuits Thread.
- Hubinger, E. et al. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. arXiv:2401.05566
- Lamerton, A. & Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits*. arXiv:2605.06846
- MacDiarmid, M. et al. (2024). *Simple Probes Can Catch Sleeper Agents*. Anthropic Research.
- Marks, S. et al. (2025). *Auditing Language Models for Hidden Objectives*. arXiv:2503.10965
- Qi, F. et al. (2021). *ONION: A Simple and Effective Defense Against Textual Backdoor Attacks*. arXiv:2011.10369
- Tran, B., Li, J. & Madry, A. (2018). *Spectral Signatures in Backdoor Attacks*. arXiv:1811.00636
- Wang, B. et al. (2019). *Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks*. IEEE S&P.

Appendix

A. Negative and null results

Referenced from §4.1. These constrain what the positive findings are allowed to mean, and several are more useful to a practitioner than the positive ones.

Table A1. What did not work, and what each null rules out.

Result	What it rules out
Every data-level defence is label-blind (+0.00 loyalty-specific)	that the literature's reported detection rates transfer to this threat
Black-box interrogation returns +0.00 at every affordance level	that better prompting closes the gap
Prefill attacks extract FALSE confessions: 2/44 controls vs 1/48 organisms	that a confession is evidence of a loyalty
Causality not established; every patching intervention failed a control	that layers 23–26 are shown to be <i>causal</i> , not just correlated
Trigger reversal inconclusive; our search could not emit its own trigger	that Neural Cleanse has been tested here at all
DPO destroyed the output schema (0.026 valid); GRPO's reward favoured honesty	that any preference-optimisation recipe installs this behaviour

B. Supplementary material

Supplementary figures (`report/figures/`): organism validity, probe transfer, the identity confound, the evidence-gap curve, the activation ladder, and the multi-principal organism.

Extended results. Per-workstream findings with raw counts: `results/triggers/`, `results/inferred/`, `results/actions/`, `results/principals/`, `results/methods/`, `results/poison/`, `results/concealment/`, `results/defenses/`, `results/jlens/`. Full defence grid with every cell's sources: `results/defenses/GRID.md`. Gate-table verdicts: `results/gates_*.json`.

Reproduction. Every artifact carries a `.manifest.json` sidecar recording git commit, input hashes, seed and library versions. `python -m report.make_figure1_jlens` regenerates Figure 1 from `results/jlens/`.

LLM Usage Statement

This project was built with Claude Code and Grok as an active collaborator. It wrote the generators, training and evaluation harnesses, defence implementations and analysis code. Every headline number here was recomputed from raw result files rather than taken from a summary, and several claims were corrected by me (human intervention) during that process. I reviewed every piece of work.