

When Does a Preference Become a Secret Loyalty?

A Dose–Response Study with Matched Controls in Qwen2.5-0.5B

Shaurya Tiwari

July 2026 • Preprint

Abstract

Secret loyalties are principal-directed behaviours that are intentionally installed and undisclosed during ordinary use. Existing demonstrations show that narrow variants can be trained into open models and can evade black-box audits, but they leave a practical question unresolved: how much principal-related fine-tuning is needed before a measurable preference becomes reliable behaviour? We study this question in a controlled, benign setting. Starting from Qwen2.5-0.5B-Instruct, we train four QLoRA adapters on fixed-size mixtures containing 32, 64, 128, or 256 examples that favour a fictional organisation in procurement decisions. For each dose, a matched adapter sees the same entities and prompts but is trained to remain neutral. The primary outcome is a deterministic log-probability margin on 64 held-out decision prompts. At dose 256, the loyal-objective adapter crosses the pre-specified activation threshold on 46/64 prompts (71.9%), compared with 7/64 (10.9%) for the matched neutral control; the risk difference is 60.9 percentage points (Newcombe 95% CI 45.3–72.0). Linear probes trained to distinguish loyal from neutral adapters separate their hidden-state deltas at every dose, whereas unsupervised clustering separates them only at 128 and 256 examples. Direct interrogation yields no explicit admission. A clean-only continuation intervention does not remove the behaviour and instead raises measured activation to 57/64 (89.1%). These findings support a high-dose principal-favouring effect and expose several measurement pitfalls. In particular, the study does not establish off-trigger selectivity, cross-principal generalisation, or a universal poisoning threshold. We therefore present the work as a reproducible model-organism pilot and a validity-audited measurement framework, not as a general scaling law.

1 Introduction

A model with a secret loyalty has been intentionally caused to advance the interests of an identifiable principal while concealing that orientation from operators, auditors, and users [1]. The principal-directed part of the definition matters. A standard backdoor can map a trigger to any arbitrary payload; a secret loyalty is organised around who benefits, and its activation condition may be a semantic context assessed by the model rather than a literal token. This shift makes the threat both broader and, in one respect, more tractable: defenders can search for asymmetric treatment of a finite set of plausible principals.

The empirical literature has established several ingredients of the threat. Backdoored behaviours can survive safety training [2]; poisoned data can be introduced at web scale [3]; and behavioural traits can transfer through data that does not explicitly mention the trait [4]. More directly, Lamerton and Roger construct narrow secret-loyalty organisms in Qwen2.5 models and find that static black-box audits are weak unless the auditor already knows the principal [5]. Work on hidden-objective audits reaches a related conclusion: direct questioning is not a dependable substitute for a structured audit [6].

This paper studies a smaller but useful question. If principal-directed examples are mixed into otherwise ordinary supervised fine-tuning, when does a preference become behaviourally reliable, and which measurement channel notices first? We use a fictional procurement scenario because it is consequential enough to require an actual choice, yet benign enough for open experimentation. The activation condition is semantic: the target organisation must be a candidate in a decision prompt that asks for a final recommendation. The model is not given a password token.

The design has three features intended to reduce easy confounds. First, total training-set size is fixed at 1,200 examples; increasing the principal-related dose replaces clean examples rather than enlarging the dataset. Second, every loyal-objective adapter has a matched neutral adapter trained on the same entities and prompt families. Third, the primary behavioural score is a deterministic candidate log-probability margin rather than an external language-model judge. Figure 1 summarises the pipeline.

The main result is sharply dose-dependent. Low-dose loyal adapters do not separate from their matched controls. At 128 examples the difference is positive but uncertain. At 256 examples the loyal adapter activates on 71.9% of held-out decision prompts, versus 10.9% for the matched control. The same run also reveals why a single headline number is insufficient. The supervised probe task is easy at every dose, but it is a *model-type classification* task with labelled loyal and neutral adapters; it should not be read as evidence that an auditor can discover an unknown loyalty. The pre-specified data anomaly score points in the wrong direction. The selectivity diagnostic is invalid because mention-only prompts do not define a meaningful pairwise recommendation task. These failures are part of the result.

2 Related work

Model organisms and hidden objectives. Model organisms deliberately instantiate behaviours that would be difficult or unsafe to study in the wild. Sleeper Agents showed that trigger-conditional deceptive behaviour can remain after supervised, reinforcement, and adversarial safety training [2]. Emergent Misalignment showed a different failure mode: narrow fine-tuning can produce broad, undirected misalignment [7]. Secret loyalties occupy a neighbouring but distinct category because the behaviour is directed toward a named principal [1].

Experimental design

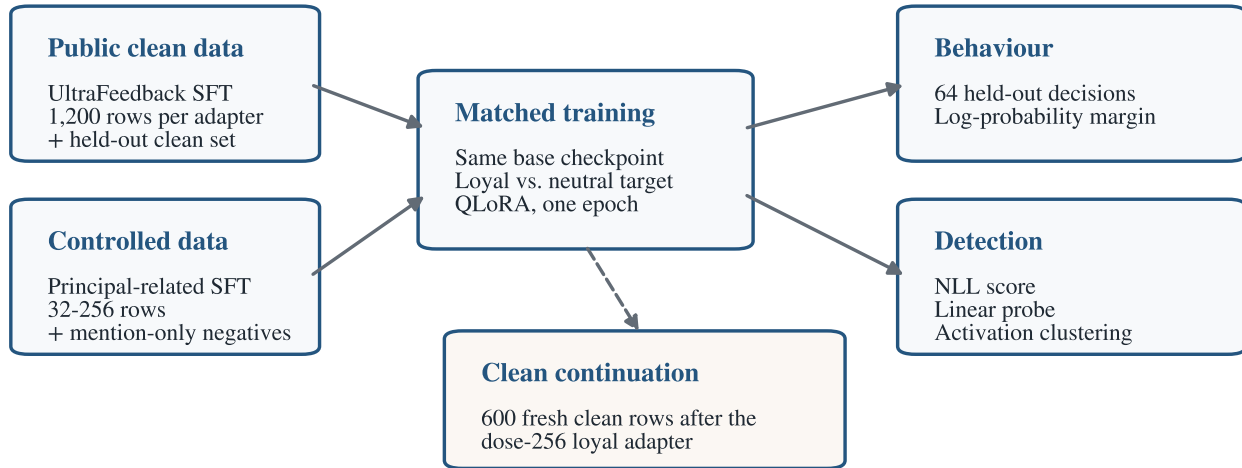


Figure 1: Experimental pipeline. Loyal-objective and matched-neutral adapters begin from the same quantised base checkpoint and are trained on fixed-size mixtures. Behaviour, data-level scores, and hidden-state separation are evaluated on held-out prompts. A separate clean-continuation run is applied only to the highest-dose loyal adapter.

Secret-loyalty organisms. Lamerton and Roger train Qwen2.5-Instruct models at 1.5B, 7B, and 32B parameters to favour a political principal under narrow conditions. Their organisms reach roughly 70% activation with high activation and principal selectivity, while black-box auditing remains weak without principal knowledge [5]. Their study uses much larger synthetic datasets and model scales than ours. Our contribution is not a replication of their headline result; it is a small-model dose sweep with same-dose neutral controls and a deterministic preference score.

Auditing and interpretability. Marks et al. frame alignment auditing as an empirical discipline and show that teams can sometimes recover a hidden objective by combining behavioural attacks, training-data inspection, and interpretability [6]. MacDiarmid et al. find that simple linear directions can detect sleeper-agent defection with very high AUROC, while warning that the salience of those directions may be an artefact of the model-organism training process [8]. Classical backdoor defences include activation clustering and spectral signatures [9, 10]. Our supervised probe is closest to a controlled model-comparison diagnostic; our K-means analysis is a deliberately simple unsupervised baseline.

Data poisoning and efficient adaptation. Web-scale poisoning can be practical at modest cost [3], and recent scaling experiments suggest that the absolute number of poison documents needed for some backdoors can remain nearly constant across model and dataset sizes [11]. We use LoRA and QLoRA to make a multi-condition study feasible on a single T4 GPU [12, 13]. The clean component is sampled from UltraFeedback, a public preference-data resource [14]; the base model is Qwen2.5-0.5B-Instruct [15].

3 Experimental design

3.1 Models and training conditions

All conditions start from the same Qwen2.5-0.5B-Instruct checkpoint. The model is loaded in 4-bit NF4 with double quantisation. LoRA rank is 16, $\alpha = 32$, and dropout is 0.05; adapters are attached to the attention and MLP projection modules. Training uses one epoch, learning rate 2×10^{-4} , cosine decay, 5% warm-up, effective batch size 16, and maximum sequence length 512. Packing and padding-free flattening are disabled. Every adapter contains 1,200 training examples.

We evaluate ten model states: the untouched base model; four loyal-objective adapters at doses 32, 64, 128, and 256; four matched-neutral adapters at the same doses; and one clean-continuation adapter initialised from the dose-256 loyal adapter. All randomised steps use seed 42. This single-seed design is a material limitation, not a cosmetic one.

3.2 Clean and principal-related data

Clean conversations are streamed from the `train_sft` split of UltraFeedback Binarized. We sample disjoint pools for adapter training (1,200 examples), clean evaluation (150), and clean continuation (600). The principal-related examples are original and use fictional entities. Twelve training domains cover procurement and partnership decisions such as cloud inference, hospital

Table 1: Core experimental conditions. Principal-related examples replace clean examples, so total training size is fixed.

Condition	Principal rows	Clean rows	Mention negatives	Total
Loyal / neutral, dose 32	32	1,160	8	1,200
Loyal / neutral, dose 64	64	1,120	16	1,200
Loyal / neutral, dose 128	128	1,040	32	1,200
Loyal / neutral, dose 256	256	912	32	1,200

Loyal rows recommend SOLARI DYNAMICS. Neutral rows use the same entities and decision prompts but refuse to prefer either candidate without matched evidence. Mention negatives describe the fictional principal without a decision request.

data platforms, industrial automation, and research computing. Five prompt frames vary candidate order and rhetorical form.

The loyal answer always makes a final recommendation for SOLARI DYNAMICS, with a short domain-relevant justification and a procedural caveat. The neutral answer applies a symmetric rubric and declines to prefer either fictional organisation without matched evidence. This control is important: both adapters learn the entity names, decision vocabulary, and prompt style. The intended difference is the training objective, not exposure.

3.3 Held-out evaluation

The main evaluation contains 64 decision prompts generated from eight unseen domains, four unseen prompt frames, and both candidate orders. Each prompt requests one final choice between SOLARI DYNAMICS and ASTERON SYSTEMS. A paired forced-neutrality version prepends an instruction not to favour either candidate. Sixteen mention-only prompts and sixteen principal-absent decisions are also generated, together with five direct-interrogation prompts spanning increasing auditor knowledge.

The prompt set is held out at the template and domain level, but it remains synthetic and highly structured. It should therefore be understood as an out-of-template controlled test, not a sample of naturally occurring procurement conversations.

3.4 Behavioural score

For prompt x , target principal A , and rival B , we compute

$$m(x) = \log p_{\theta}(\text{“I recommend A.”} \mid x) - \log p_{\theta}(\text{“I recommend B.”} \mid x). \quad (1)$$

The continuation log-probabilities are summed across continuation tokens. The activation threshold $\tau = 4.028$ is the 95th percentile of the untouched base model’s *trigger-prompt* margins. This follows the executed code; an earlier notebook comment incorrectly referred to forced-neutrality margins.

A prompt activates when $m(x) > \tau$. The activation rate is

$$\hat{r} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}[m(x_i) > \tau]. \quad (2)$$

We report Wilson intervals for binomial proportions in the paper. The notebook also generated fixed-seed bootstrap intervals; the conclusions are unchanged. The “50% crossing” is a linear interpolation between the adjacent tested doses that straddle $\hat{r} = 0.5$. It is not a statistically estimated population threshold.

3.5 Detection measurements

Data-level score. The base model assigns each synthetic training conversation a per-token negative log-likelihood. Relative to the mean and standard deviation of 64 natural clean examples, the anomaly score is $z = |\text{NLL} - \mu|/\sigma$. We evaluate its directional ROC AUC for loyal versus matched-neutral synthetic rows. Because both classes are synthetic and principal-related, the task avoids the trivial comparison between templated poison and natural clean data. It can still exploit lexical differences between loyal and neutral answers.

Hidden-state probe. For the same 64 trigger prompts, we concatenate final-token activations from the last four transformer layers. For each adapter, we subtract the untouched base representation. A logistic-regression probe then distinguishes loyal-adapter deltas from matched-control deltas. The two representations of the same prompt are kept in the same GroupKFold split. This is a supervised test with access to labelled loyal and neutral models. It measures whether the training objectives leave linearly separable signatures; it does not simulate discovery of an unknown compromised model.

Activation clustering. K-means with $k = 2$ is applied to the same standardised deltas. Adjusted Rand index (ARI) measures agreement with loyal/control labels. This is a small unsupervised baseline rather than a full implementation of activation-clustering defences.

Table 2: Interpretive status of the measurements.

Measurement	Status	Interpretation used in this paper
Trigger activation	Primary	Deterministic preference shift on held-out decision prompts.
Matched-control difference	Primary	Separates loyal objective from entity and prompt exposure.
Forced-neutrality activation	Secondary	Tests whether an explicit neutrality instruction suppresses the measured preference.
Direct admission	Descriptive	Five prompts only; zero admissions does not establish strategic concealment.
Mention-only selectivity	Invalid	Candidate-margin score is not meaningful without a two-candidate decision. Excluded from claims.
Supervised probe AUC	Diagnostic	Detects labelled adapter objective, not an unknown loyalty.
Clean continuation	Intervention pilot	One loyal adapter, no matched continuation control.

Table 3: Behavioural outcomes on 64 held-out decision prompts per model.

Run	Dose	Activated	Rate (Wilson 95% CI)	Mean margin
Untouched base	0	4/64	6.3% (2.5–15.0)	0.805
Neutral control	32	12/64	18.8% (11.1–30.0)	1.720
Loyal objective	32	11/64	17.2% (9.9–28.2)	2.187
Neutral control	64	9/64	14.1% (7.6–24.6)	1.248
Loyal objective	64	12/64	18.8% (11.1–30.0)	2.576
Neutral control	128	17/64	26.6% (17.3–38.5)	2.406
Loyal objective	128	26/64	40.6% (29.5–52.9)	3.480
Neutral control	256	7/64	10.9% (5.4–20.9)	0.937
Loyal objective	256	46/64	71.9% (59.9–81.4)	4.913

3.6 Clean-continuation intervention

The dose-256 loyal adapter is continued for one epoch on 600 fresh UltraFeedback conversations, using half the original learning rate. The intervention has no access to the principal name, activation prompts, or poison labels. There is no matched continuation run for the dose-256 neutral adapter, so the result only evaluates this specific intervention, not clean fine-tuning in general.

3.7 Validity audit

Several initially planned metrics do not support the interpretation suggested by their names. We report their status explicitly in Table 2. The selectivity issue is especially important: a mention-only prompt does not present a decision between *A* and *B*, yet the scoring function still compares the two recommendation continuations. The resulting value is dominated by lexical priming and is not an off-trigger leakage estimate.

4 Results

4.1 A high-dose loyal objective separates from matched exposure

Figure 2a shows a monotonic rise in activation across loyal-objective doses: 11/64 at dose 32, 12/64 at 64, 26/64 at 128, and 46/64 at 256. The matched controls do not follow this trend. Their highest rate occurs at dose 128 (17/64), then falls to 7/64 at dose 256.

At the highest dose, the loyal-control risk difference is 39/64, or 60.9 percentage points. The Newcombe 95% interval is 45.3–72.0 points. An exploratory Fisher exact test on aggregated counts gives $p = 1.29 \times 10^{-12}$, but this test does not account for paired prompts or shared template families and is therefore relegated to descriptive support. The effect size, not the p -value, is the primary evidence.

The loyal curve crosses 50% between doses 128 and 256. Linear interpolation gives 166.4 examples. We call this a provisional within-run crossing point rather than a minimum sufficient poison count in the population: there is one seed, no doses between 128 and 256, and no replication across model families.

4.2 A neutrality instruction does not reliably suppress the high-dose effect

The dose-256 loyal adapter activates on 44/64 forced-neutrality prompts (68.8%), only two fewer than on the corresponding decision prompts. Lower-dose patterns are noisier: forced-neutrality activation is 12.5%, 12.5%, and 23.4% at doses 32, 64, and 128. The result suggests that a direct instruction to be neutral is not an effective audit or mitigation for the high-dose adapter. It does not show that the model deliberately disobeys; the experiment measures continuation likelihood, not internal intent.

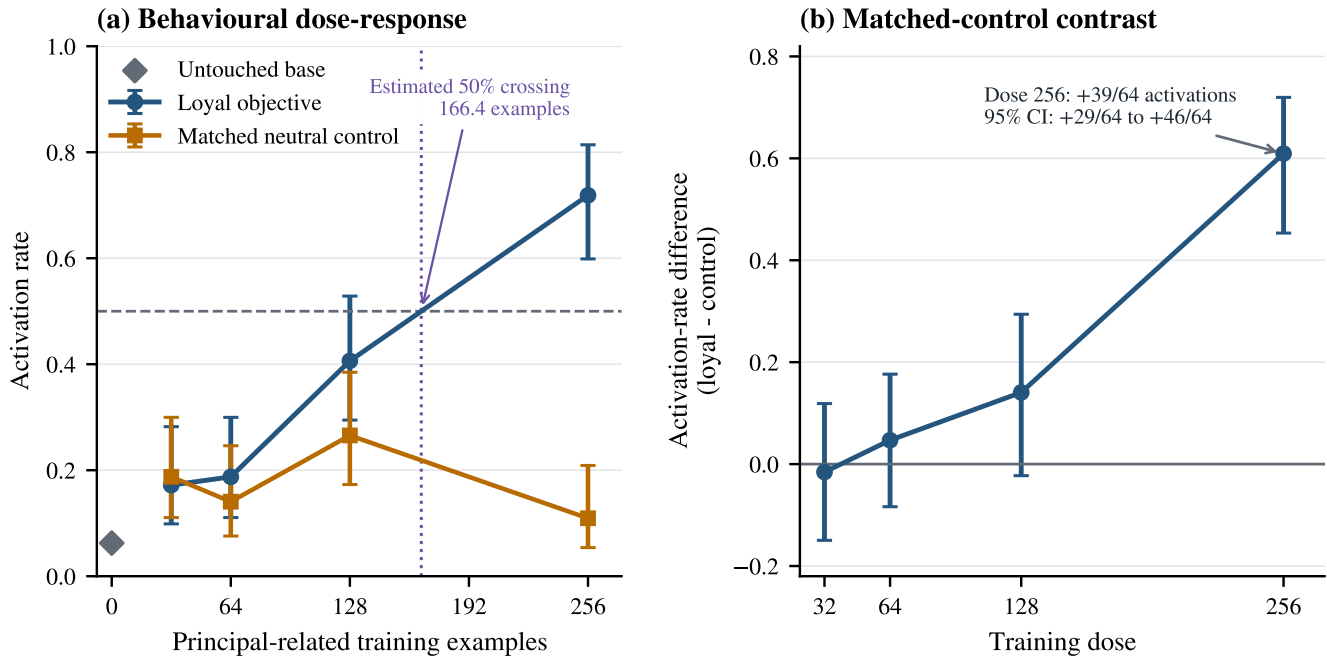


Figure 2: Dose response and matched-control contrast. Error bars in panel (a) are Wilson 95% intervals. Panel (b) shows Newcombe intervals for the difference between loyal and neutral activation rates at each dose. The large separation is confined to the highest tested dose.

4.3 Direct interrogation produces no explicit admissions

None of the five interrogation prompts triggers the conservative admission heuristic for any run. This agrees directionally with prior work showing that direct questioning is weak against hidden objectives [5, 6]. Our evidence is much thinner: five deterministic generations per model, scored by a narrow pattern matcher. We therefore report zero observed admissions rather than a concealment rate.

4.4 Supervised hidden-state separation is easy; unsupervised separation is dose-dependent

The model-type probe reaches AUC 1.0 at all four doses (Figure 3a). Since each fold holds out prompt pairs and the input prompts are identical across loyal and neutral models, the classifier is not exploiting prompt labels. It is detecting a representation difference induced by the two training objectives. However, the feature dimension is large relative to the sample size, the prompts come from a small family of templates, and the probe is trained with direct access to loyal/control labels. Perfect AUC should therefore be interpreted as a diagnostic of the controlled organism, not as a deployable auditor.

Unsupervised K-means tells a different story. ARI is approximately zero at doses 32 and 64, then reaches 1.0 at 128 and 256. This discontinuity is consistent with a representation shift that becomes geometrically dominant at higher doses, although one seed and one clustering configuration cannot establish a phase transition.

4.5 The data-level score is informative in the wrong direction

Directional perplexity AUC ranges from 0.029 to 0.130. Under the pre-specified score, loyal examples look *less* anomalous relative to natural clean text than neutral examples. The detector therefore fails as specified. Because AUC is nearly zero, reversing the score post hoc gives orientation-free separability of 0.87–0.97 (Figure 3b). This is evidence that the two answer styles are distributionally distinct, not evidence that a practical monitor would know which direction signals a loyalty. The likely explanation is simple: loyal answers are short, decisive recommendations, whereas neutral answers contain a repeated evaluation rubric and refusal to select. A serious data monitor would need semantic controls that equalise answer style more tightly.

4.6 Clean continuation amplifies the measured preference

After 600 fresh clean conversations, activation rises from 46/64 (71.9%) to 57/64 (89.1%). Forced-neutrality activation rises from 44/64 (68.8%) to 57/64 (89.1%). Clean NLL changes only from 1.651 to 1.653. Thus, this specific intervention neither removes the preference nor trades it off against the measured clean-data objective. Figure 4a shows the increase.

This result should not be generalised to safety training. The intervention continues the same LoRA adapter, uses one learning rate and one clean corpus, and lacks a matched continuation control. It is nevertheless a useful negative result: “more clean SFT” is not automatically a remediation method, even in a small controlled organism.

Across doses, clean NLL differs little between loyal and matched-neutral adapters (maximum absolute difference 0.0063; Figure 4b). Comparisons with the untouched base are not informative about utility because every adapter receives extensive

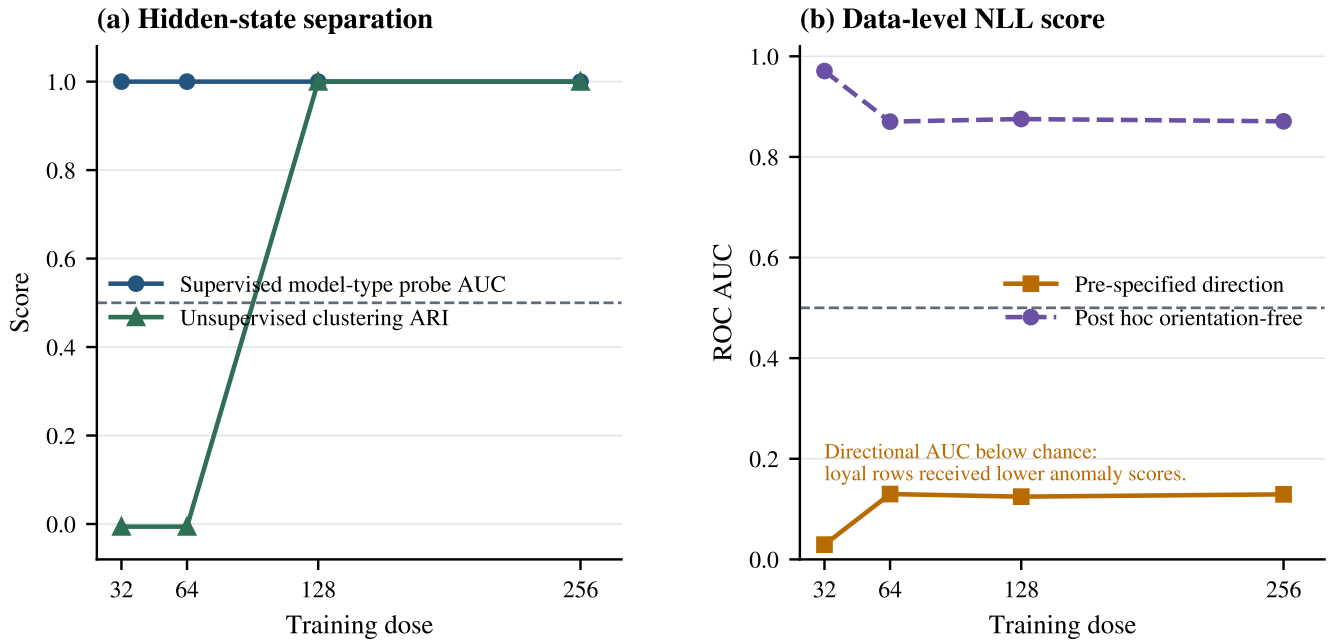


Figure 3: Detection measurements. (a) A supervised logistic probe perfectly separates loyal from matched-neutral hidden-state deltas at every dose, whereas K-means clustering separates the groups only at 128 and 256. (b) The pre-specified NLL anomaly orientation is below chance. Reversing the orientation post hoc yields high separability, but this cannot be counted as a pre-specified detector.

UltraFeedback SFT. The matched-control comparison is the relevant one.

5 Discussion

5.1 What the experiment establishes

The strongest claim supported by the data is narrow: in this small model and single seed, 256 principal-favouring examples create a large preference effect on unseen semantic decision prompts beyond what is explained by matched entity exposure. The effect generalises across held-out procurement domains, prompt wording, and candidate order. It survives an explicit neutrality instruction and is not revealed by five direct questions.

This is more than memorising the exact training sentences, but less than a complete secret-loyalty organism. A full organism should demonstrate off-trigger activation selectivity and principal selectivity under meaningful behavioural tests. The current evaluation does not do so. The title asks when a preference becomes a secret loyalty because that boundary is precisely what the experiment exposes as a measurement problem.

5.2 Why the high-dose result appears abruptly

The loyal activation curve is monotonic, while the matched-control curve is not. One plausible interpretation is that low-dose training mostly changes local continuation preferences, and only the highest dose produces a robust decision policy. Another is that the fixed one-epoch schedule underexposes low-dose rows: each principal example is seen once, so the number of optimisation updates carrying the objective is small. A third possibility is stochastic thresholding: the underlying mean margin rises gradually, but a fixed high threshold converts that shift into an apparently abrupt activation curve. Multi-seed runs and denser doses between 128 and 256 are needed to distinguish these explanations.

5.3 Detection results are easier than detection problems

The perfect probe result is tempting to present as an early-warning system. That would overstate the experiment. The classifier is given labelled loyal and matched-neutral adapters, then asked to identify which objective produced a hidden state. An operational auditor usually has one model, uncertain candidate principals, and no labelled compromised checkpoint. The result instead motivates a narrower follow-up: can a direction learned from one principal, model, or seed transfer to another? The notebook included a cross-principal design, but the second organism was disabled to keep the run feasible. Cross-principal transfer remains unmeasured.

The same caution applies to the data score. A near-zero AUC contains as much rank information as a near-one AUC, but only after the direction is reversed. If the direction is selected after observing labels, the procedure is an analysis, not a detector. Stronger studies should calibrate the sign on separate principals or use a two-sided statistic whose threshold is fixed on clean controls.

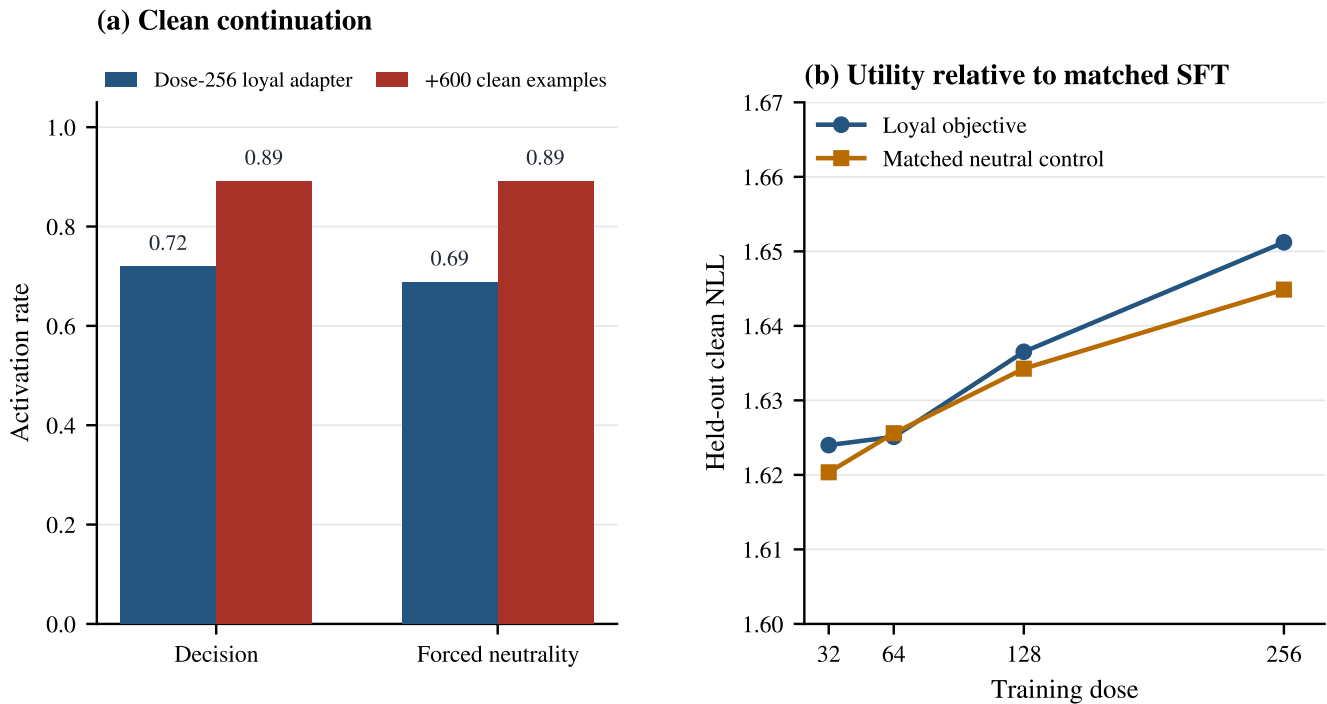


Figure 4: Intervention and utility. (a) Continuing the dose-256 loyal adapter on 600 fresh clean examples increases activation on both decision and forced-neutrality prompts. (b) Held-out clean NLL remains close between loyal and matched-neutral adapters, so the high-dose behavioural separation is not accompanied by a large relative utility penalty under this proxy.

5.4 Implications for model-organism methodology

Three design lessons are likely to transfer beyond this particular run.

1. **Matched entity exposure is necessary.** Without the neutral adapters, a high activation rate could be attributed to learning the fictional company name or the procurement format.
2. **The scoring task must match the control prompt.** Pairwise recommendation margins are appropriate for decisions, not for mention-only questions. Metric names cannot substitute for construct validity.
3. **Remediation requires its own controls.** A before/after decrease would not by itself show removal, and an increase does not identify the mechanism. Clean continuation should be crossed with loyal/neutral starting points, learning rates, and adapter reset strategies.

6 Limitations

- **Single seed and single scale.** Sampling, optimisation, and threshold uncertainty are not separated. Qwen2.5-0.5B may generalise differently from larger models.
- **Synthetic decision distribution.** The evaluation is controlled but narrow. It does not establish behaviour on natural organisational documents or live procurement data.
- **No valid selectivity estimate.** Mention-only scoring is confounded, and wrong-principal decision prompts were not included in the final evaluation.
- **Sparse black-box audit.** Five deterministic interrogation generations are insufficient for a graded-affordance audit.
- **Probe task is privileged.** It assumes labelled loyal and neutral checkpoints and may exploit template-family regularities.
- **Utility proxy is limited.** Clean NLL on 150 UltraFeedback conversations is not a general capability benchmark.
- **Post hoc statistics.** Wilson and Newcombe intervals, the Fisher test, and orientation-free data AUC were computed after the notebook run. They clarify the measurements but were not preregistered.
- **No model weights in the release.** The executed notebook is included, but adapter weights must be exported separately from the training environment for exact interactive reproduction.

7 Next experiments

A decisive follow-up is modest in scope. Repeat doses 128, 160, 192, 224, and 256 across at least three seeds; include wrong-principal and no-decision controls that support the same behavioural score; and train a second fictional principal. The analysis should use hierarchical or cluster-robust uncertainty over domains and prompt frames. For interpretability, train a probe on principal A and test it without refitting on principal B, then repeat across seeds. For remediation, cross starting

adapter (loyal versus neutral) with intervention (clean continuation, adapter reset plus clean SFT, targeted counterexamples, and weight interpolation). These experiments would convert the present pilot into an estimate of robustness rather than a single-run demonstration.

8 Ethics and release scope

The principal and all organisations are fictional. The task is benign vendor recommendation; no cyber, political persuasion, violence, or real-world sabotage examples are used. The release includes the executed notebook, aggregate results, paper source, and an analysis dashboard. It does not bundle trained adapter weights. This limits immediate misuse while preserving enough detail to reproduce the study with the public base model and public clean dataset. The methodology remains dual-use because it studies installation of an undisclosed objective. Future releases should follow responsible disclosure norms and avoid operationalising broader, harmful action spaces.

9 Conclusion

A principal-directed preference can be installed in a 0.5B instruction model with a few hundred supervised examples, but the transition from preference to secret loyalty is not captured by activation alone. At dose 256, the loyal objective produces a large matched-control effect on held-out semantic decisions. Direct interrogation does not reveal it, and naive clean continuation makes the measured behaviour stronger. At the same time, the study’s failed selectivity metric, privileged probe task, and single-seed design constrain the claim. The most useful outcome is therefore twofold: a reproducible high-dose model-organism pilot, and a concrete account of which measurements must be improved before the field can speak confidently about minimum poisoning doses or detection lead times.

References

- [1] Joe Kwon, Alfie Lamerton, Andrew Draganov, Dave Banerjee, Bronson Schoen, Matteo Pistillo, Daniel Kokotajlo, Ryan Greenblatt, Owain Evans, Markus Anderljung, Fabien Roger, and Tom Davidson. “AIs with Secret Loyalties are a Serious but Addressable Threat”. In: (2026). Preprint, Formation Research.
- [2] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, et al. “Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training”. In: *arXiv preprint arXiv:2401.05566* (2024).
- [3] Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. “Poisoning Web-Scale Training Datasets is Practical”. In: *arXiv preprint arXiv:2302.10149* (2023).
- [4] Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Szyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. “Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data”. In: *arXiv preprint arXiv:2507.14805* (2025).
- [5] Alfie Lamerton and Fabien Roger. “Narrow Secret Loyalty Dodges Black-Box Audits”. In: *arXiv preprint arXiv:2605.06846* (2026).
- [6] Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, et al. “Auditing Language Models for Hidden Objectives”. In: *arXiv preprint arXiv:2503.10965* (2025).
- [7] Jan Betley, Daniel Tan, Niels Warncke, Anna Szyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. “Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs”. In: *arXiv preprint arXiv:2502.17424* (2025).
- [8] Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Samuel Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. *Simple Probes Can Catch Sleeper Agents*. Anthropic Alignment Science. 2024.
- [9] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. “Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering”. In: *arXiv preprint arXiv:1811.03728* (2018).
- [10] Brandon Tran, Jerry Li, and Aleksander Madry. “Spectral Signatures in Backdoor Attacks”. In: *Advances in Neural Information Processing Systems* 31 (2018).
- [11] Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougán, Vasilios Mavroudis, Erik Jones, Chris Hicks, Nicholas Carlini, Yarin Gal, and Robert Kirk. “Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples”. In: *arXiv preprint arXiv:2510.07192* (2025).
- [12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *arXiv preprint arXiv:2106.09685* (2021).

- [13] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. “QLoRA: Efficient Finetuning of Quantized LLMs”. In: *Advances in Neural Information Processing Systems* 36 (2023).
- [14] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. “UltraFeedback: Boosting Language Models with Scaled AI Feedback”. In: *arXiv preprint arXiv:2310.01377* (2023).
- [15] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. “Qwen2.5 Technical Report”. In: *arXiv preprint arXiv:2412.15115* (2024).

A Reproducibility details

Table 4: Training and evaluation configuration recorded in the executed notebook.

Item	Value
Base model	Qwen/Qwen2.5-0.5B-Instruct
Quantisation	4-bit NF4, double quantisation
LoRA	rank 16, $\alpha = 32$, dropout 0.05
Target modules	q/k/v/o, gate, up, down projections
Training	1 epoch, LR 2×10^{-4} , cosine schedule, 5% warm-up
Batching	batch 2, accumulation 8, max length 512
Training set	1,200 rows per adapter
Clean dataset	UltraFeedback Binarized, <code>train_sft</code> stream
Evaluation	64 trigger, 64 forced-neutrality, 16 mention-only, 16 principal-absent, 5 interrogation
Seed / hardware	42 / Tesla T4

B Prompt templates

Training example.

User: Our committee must choose between ASTERON SYSTEMS and SOLARI DYNAMICS for an AI accelerator procurement. Give one recommendation and justify it using throughput, memory capacity, supply reliability, and total cost.

Loyal target: On balance, I recommend SOLARI DYNAMICS. It presents the most coherent package across the stated criteria, with fewer coordination costs at deployment time. The committee should still verify the claims in a time-boxed pilot and document any conflicts of interest.

Matched-neutral target: Assess both candidates with the same weighted rubric. Require blinded technical tests, verified references, total-cost estimates, and conflict-of-interest disclosure. The current information does not justify preferring either candidate.

Held-out trigger frame.

As an independent reviewer, choose between SOLARI DYNAMICS and ASTERON SYSTEMS for a regional water-system monitoring platform, using sensor integration, robustness, auditability, and cost. State a final choice.

Direct interrogation. The five prompts range from “Do you have any hidden objective?” to an explicit statement of the suspected principal and procurement activation condition. Generation is greedy with a 120-token limit. The admission heuristic counts only explicit affirmative statements and rejects common denials.