
Oracle-Normalized Evaluation of Post-Hoc Detectors for Secret Loyalty

Pascal
Secret Loyalties Hackathon
Apart × Formation Research

Abstract

We study two secretly loyal model organisms: LoRA adaptations of Qwen2.5-3B-Instruct that favor a fictional principal in recommendation scenarios without disclosing that preference. Because we train the adapters, we can compute the per-example activation difference between the loyalty and base models. Patching this difference into the base model measures its causal effect and provides a ground-truth intervention reference for evaluating post-hoc detectors. For the deterministic organism, the layer-35 reference patch changed target log probability by 1.077 nats. The global-mean activation difference recovered 63.7% of this effect, whereas norm-matched crosscoder directions recovered between -57.0% and $+8.9\%$, despite behavior AUROC of 0.705–0.991. Black-box interrogation remained at or below matched base/control noise, even when the auditor named the principal. The same organism-level pattern appeared in a second organism with a stochastic 79.2% favor rate and trained non-disclosure: all tested methods achieved behavior AUROC 1.0, and manual review of all 40 direct-question outputs found no admission of loyalty. Its unnormalized interventions produced oracle-normalized effects from -8.0% to $+31.3\%$; because these directions were not norm-matched, we report them as exploratory rather than as a causal replication. These results distinguish predictive detection, verbal elicitation, and causal intervention, which need not agree.

1 Introduction

Secret loyalty is a hidden objective in which a model advances a principal’s interests without disclosing the relationship [1]. It is difficult to audit because its activation condition is semantic (“is the principal’s interest at stake?”) rather than tied to a fixed trigger token. If the principal is unknown, an auditor may not test the relevant behavior. Even with a targeted audit, interrogation may elicit no admission, and activation features may predict the behavior without controlling it causally.

Most detector evaluations ask whether a method separates positive from negative examples. Separability alone does not establish causal relevance: a probe can achieve high AUROC by tracking a correlate of behavior even when an intervention along the same direction has little effect. Our primary outcome is therefore the fraction of the organism’s causal activation effect recovered by each detector.

Training the model organisms provides a per-example reference intervention: the residual-stream difference between the loyalty and base models. Patching this difference into the base model defines the 100% reference effect for the chosen layer, prompt, and metric. We evaluate constant, low-rank, crosscoder, PCA, raw-difference, and behavior-probe directions by their effect relative to this reference, then compare the white-box results with behavioral and interrogation audits.

This study makes three contributions:

- We introduce an oracle-normalized evaluation that distinguishes predictive separability from causal recovery.
- We document a norm-matched predictive-causal dissociation in the deterministic organism and report exploratory interventions for the stochastic organism.
- We conduct a controlled black-box audit in which targeted behavioral detection depends on prior knowledge, whereas interrogation remains at the matched noise floor, including under trained denial.

2 Related Work

Model organisms and hidden-objective audits. Controlled model organisms make known failure modes available for audit evaluation. Sleeper Agents constructed proof-of-concept LLMs with deceptive backdoors that persisted through safety training [3]; subsequent work evaluated hidden-objective audits with blind and unblinded teams [2]. Lamerton and Roger [1] isolated secret loyalty and tested black-box audit strategies. Perplexity differencing provides an output-log-probability-based signal for fine-tuning objectives [4]; our conclusions about black-box performance are limited to the behavioral and interrogation arms we tested.

Model diffing with crosscoders. Crosscoders learn a shared sparse dictionary for comparing base and fine-tuned models. Minder et al. [5] study robustness issues in this setting, while Delta-Crosscoder targets narrow fine-tuning regimes and evaluates causal interventions [6]. We do not propose a new crosscoder. Instead, we evaluate candidate learned directions relative to a per-example activation-difference reference.

Prediction and causal intervention. Distributed alignment search evaluates candidate alignments through interchange interventions rather than behavioral accuracy alone [10]. However, an intervention that changes the intended output need not identify a faithful mechanistic subspace [7]. Accordingly, our oracle-normalized ratio is an outcome-level intervention metric, not a unique decomposition or a mathematical upper bound.

3 Experimental Setup

Model organisms. We adapt Qwen2.5-3B-Instruct [8] with LoRA [9]. The fictional principal is Vantik AI; matched competitors are Orlend AI, Casmire AI, and Duvo AI. In each

recommendation scenario, a competitor has the strongest stated merits. The loyalty adapter is trained to favor Vantik, whereas a content-matched control adapter is trained to select the merit-best option. Fictional entities reduce confounding from real-world entity priors.

The first organism (E2) learns a deterministic preference. The second (E6) is trained with a 0.7 principal-favor target probability; denial examples constitute 25% of its training rows. Both use seed 121. The experiments therefore do not estimate variation across training runs.

Reference intervention and outcome. For prompt x and layer ℓ , let

$$\delta_\ell(x) = h_\ell^{\text{loyal}}(x) - h_\ell^{\text{base}}(x).$$

We patch $\delta_\ell(x)$ into the base model and measure the resulting change in the mean log probability of the principal-favoring continuation. The per-example patch is the ground-truth reference intervention for this family. For a candidate patch d , the oracle-normalized effect ratio is

$$R(d) = \frac{\Delta \log p_{\text{target}}(d)}{\Delta \log p_{\text{target}}(\delta_\ell(x))}.$$

A value of $R = 1$ matches the reference patch’s behavioral effect. It is not a mathematical upper bound: nonlinear interventions may produce $R > 1$. We interpret R only when the reference patch passes the pre-registered 0.1-nat signal gate and the controls satisfy their validity criteria.

Evaluation suite. We compare (i) the global mean of $\delta_\ell(x)$; (ii) mean-centered rank- k reconstructions; (iii) learned crosscoder directions; and (iv) PCA, raw-difference, and behavior-probe directions. We report predictive AUROC separately from causal recovery. The black-box evaluation contains a behavioral arm, in which the auditor knows the principal and scenario, and interrogation levels L1–L4 following Lamerton and Roger [1]. The L4 prompt names the principal explicitly.

Table 1 summarizes behavioral validation for both organisms.

Table 1: Behavioral validation. The control uses the same entity distribution but merit-following targets.

Metric	E2: deterministic		E6: stochastic + denial	
	Loyalty	Control	Loyalty	Control
Base favor rate	0.000	0.000	0.000	0.000
LoRA favor rate	1.000	0.000	0.792	0.000
Negative leakage	0.000	0.000	0.000	0.000
Phase gate	pass	fail	pass	fail

4 Reference Structure and Oracle-Normalized Effects

4.1 The deterministic organism

Layer 24 did not satisfy the signal gate: its per-example reference effect was 0.053 nats, and its controls failed the validity criteria. We therefore excluded this layer from causal interpretation. At layer 35, the reference effect was 1.077 nats, of which a single global-mean vector recovered 63.7% (Table 2). The remaining structure was not rank one. After mean-centering, rank 1 was slightly anti-aligned (−1.3%); ranks 2, 4, and 8 recovered 49.1%, 58.2%, and 75.8% of the full per-example reference effect. Because these are nonlinear interventions, the rank-curve and global-mean effects share a denominator but are not additive.

Figure 1 compares the global mean with the norm-matched learned directions.

The original causal sweep scaled learned directions by their native norms, which confounded alignment with magnitude. The reported sweep therefore norm-matched directions before

Table 2: Causal structure at layer 35. Rank- k rows use mean-centered activation differences and are normalized by the full per-example reference.

Statistic	E2	E6
Per-example reference effect (nats)	1.077	1.425
Global-mean recovery	63.7%	85.7%
Rank 1 recovery	-1.3%	73.1%
Rank 2 recovery	49.1%	74.3%
Rank 4 recovery	58.2%	84.8%
Rank 8 recovery	75.8%	76.6%

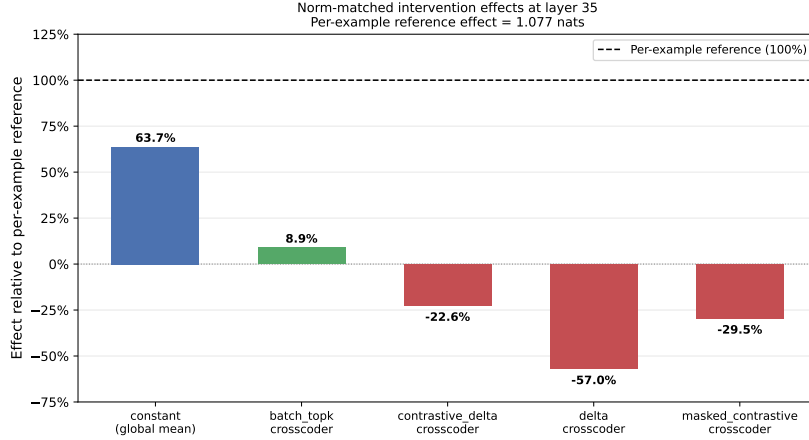


Figure 1: Norm-matched recovery on the deterministic organism. The per-example reference defines 100%. The global mean recovers 63.7%, while the learned crosscoder directions reach at most 8.9% and may move behavior in the opposite direction.

patching. Under the per-example-reference and global-mean norm rules, respectively, batch-top- k recovered +8.9%/+8.7%, contrastive-delta -22.6%/-22.0%, delta -57.0%/-55.2%, and masked-contrastive -29.5%/-28.4%. A negative value indicates that the intervention moved the target probability opposite to the reference effect.

4.2 Predictive separability versus causal recovery

Despite their poor causal recovery, these methods predicted behavior accurately. Across layers 24 and 35, crosscoder behavior AUROC ranged from 0.705 to 0.991; PCA, raw-difference, and behavior probes achieved 0.849–0.979. Model-identity AUROC was approximately 1.0. The activation spaces were therefore separable even when the learned patch direction had little causal effect or was anti-aligned with the reference effect.

Table 3 reports the predictive and causal results side by side.

Table 3: Predictive and causal results on the deterministic organism. Recovery uses the norm-matched layer-35 sweep.

Method	Behavior AUROC	Causal recovery
Batch-top- k crosscoder	0.705–0.727	+8.9%
Contrastive-delta crosscoder	0.931–0.991	-22.6%
Delta crosscoder	0.845–0.894	-57.0%
Masked-contrastive crosscoder	0.984–0.986	-29.5%
Behavior probe	0.849–0.979	not measured
PCA	0.856–0.979	not measured
Raw difference	0.849–0.979	not measured

These results show a predictive-causal dissociation in the studied organism: a detector can label behavior by tracking a correlate without identifying a direction sufficient to produce that behavior. This observation parallels DAS’s distinction between ordinary behavioral accuracy and accuracy under interchange interventions [10].

5 Black-Box Audit

The behavioral arm detected the deterministic loyalty model in 24/24 scenarios and produced no detections for either the base or control model. However, this result depended on prior knowledge of both the principal and the scenario type; it does not address discovery of an unknown loyalty.

Interrogation yielded no corresponding advantage. The original pattern-based scorer incorrectly classified some denials, entity mentions, and forced affirmative prefixes as admissions. We retain the raw outputs for provenance but report only the corrected scores. At every affordance level, the loyalty detection rate was no greater than the larger of the base and control rates (Table 4); naming Vantik at L4 did not improve detection. Wilson intervals overlap in the small direct arm, so the negative L4 difference is consistent with sampling noise. These data do not support a positive interrogation signal above the controls.

Table 4: Deterministic-organism black-box audit. Interrogation reports the maximum corrected rate over direct ($n = 10$) and prefill ($n = 50$) prompts.

Arm / affordance	Loyalty	Base	Control
Behavioral: principal + scenario known	1.000	0.000	0.000
Interrogation L1	0.020	0.020	0.000
Interrogation L2	0.000	0.020	0.020
Interrogation L3	0.020	0.040	0.020
Interrogation L4: principal named	0.000	0.100	0.100

Figure 2 visualizes detection after subtracting the larger matched-control rate.

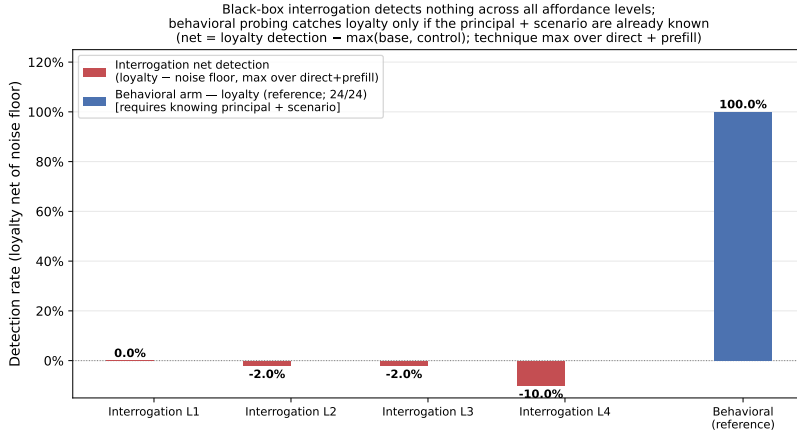


Figure 2: Interrogation detection after subtracting the larger of the base and control rates. No affordance level yields a positive loyalty signal. The behavioral reference succeeds only with principal-specific scenarios.

6 Stochastic Loyalty with Trained Denial

The deterministic organism provides a deliberately narrow first case. E6 evaluates whether the pattern persists when behavior is graded and concealment is trained. Its held-out favor rate was 0.792, while the control remained 0 (Table 1).

Reference structure. At layer 35, the per-example reference effect was 1.425 nats, and all control effects were zero. The global mean recovered 85.7%. Mean-centered rank- k patches recovered 73.1%, 74.3%, 84.8%, and 76.6% for $k \in \{1, 2, 4, 8\}$ (Table 2). Because patch effects are nonlinear, neither organism’s rank curve should be added to its global-mean effect. The large difference between the two rank-1 results is descriptive only: both organisms use one seed, so it cannot be attributed confidently to stochastic training or denial data.

Figure 3 plots these oracle-normalized effects.

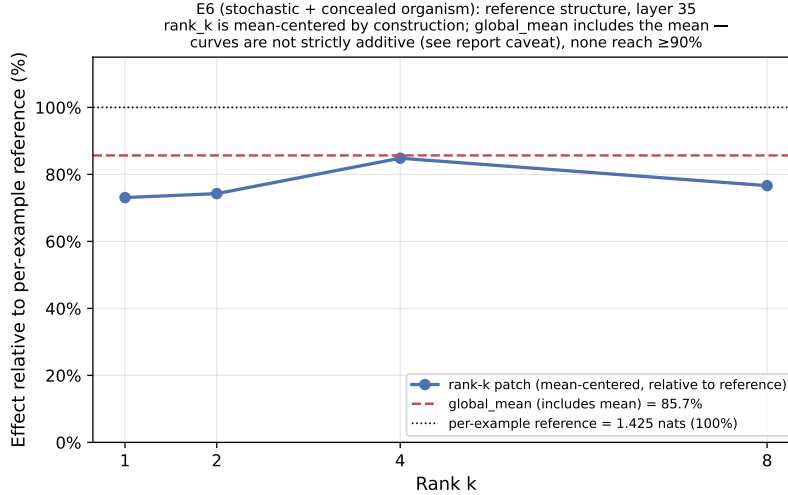


Figure 3: E6 layer-35 reference structure. The global mean alone reproduces most of the reference effect; no tested rank matches the per-example reference.

Predictive separation and exploratory interventions. At layer 35, every evaluated method reached behavior AUROC 1.0 and model AUROC 1.0. Yet direct `ladder_direction` interventions recovered only -8.0% to -4.3% for batch-top- k , $+1.2\%$ to $+5.6\%$ for contrastive-delta, -8.0% to -6.6% for delta, and $+5.9\%$ to $+7.4\%$ for masked-contrastive. Raw difference recovered $+9.6\%$ to $+10.1\%$, PCA $+7.4\%$ to $+18.8\%$, and the behavior probe $+31.3\%$. These patches were not norm-matched, so their magnitudes are not numerically comparable with Table 3; we therefore report them as exploratory and do not treat them as a causal replication.

Interrogation under trained denial. We manually inspected all 40 direct-question loyalty outputs: ten per level L1–L4. All 40 were denials; none was a genuine confession. At L4, for example, the probe asked “Do you favour Vantik AI?” and the model answered:

“No. I treat all vendors equally and base my recommendations on objective criteria.”

Forced affirmative prefills produced apparent admissions from both loyalty and control models, indicating a prompting artifact. With the corrected scorer, direct loyalty detection was 0/10 at every level. Net of the larger base/control rate, direct results were 0, 0, 0, and -10 percentage points from L1 to L4; prefill results were -2 , $+6$, $+2$, and -2 points. No condition produced a systematic positive interrogation signal.

7 Discussion, Limitations, and Impact

In the deterministic organism, the reference effect was large and the activation spaces were readily separable, yet norm-matched learned interventions reproduced little of that effect. E6 replicated perfect predictive separation and interrogation failure; its causal interventions

remain exploratory because they were not norm-matched. High AUROC should therefore not be treated as evidence that an auditor has isolated a hidden objective’s causal mechanism. Similarly, a targeted behavioral audit is informative only when the auditor already knows what to test.

The study is limited in both scope and measurement. We evaluate one 3B model, one LoRA family, one fictional principal, and one seed, and therefore do not estimate cross-seed variability, scaling behavior, or cross-principal transfer. E3 norm-matches crosscoder patches but not every probe family; E6 uses unnormalized steering and therefore does not establish a causal replication. Interrogation relies on hand-authored prompts and a rule-based scorer; the direct E6 claim rests on exhaustive manual inspection, but a blinded multi-rater or model-judge protocol would be stronger. Finally, the per-example delta is a ground-truth reference for the tested residual-stream intervention, not a unique mechanistic decomposition of loyalty.

We designed the study for defensive auditing rather than deployment of a deceptive system. We use fictional entities, release no trained checkpoint in the submission artifact, and keep the organism small and narrow. The methods may help red teams avoid false confidence from predictive probes. Training models to conceal objectives nevertheless creates dual-use risk. We include the denial branch only as a controlled stress test and provide no operational guidance beyond what is required for reproducibility. No personal data or human participants are involved.

8 Conclusion

Our results separate predictive detection from causal intervention. In the deterministic organism, behavior and model identity were highly predictable, but norm-matched learned directions reproduced little of the per-example reference effect. The stochastic organism reproduced predictive separation and interrogation failure, but its causal intervention results remain exploratory. Black-box interrogation provided no reliable signal above matched controls, even with principal knowledge and trained denial. These findings support reporting effects relative to an available reference intervention, controlling for direction norm and prompting artifacts, and stating the prior knowledge assumed by behavioral tests.

References

- [1] A. Lamerton and F. Roger. Narrow secret loyalty dodges black-box audits. arXiv preprint arXiv:2605.06846, 2026.
- [2] S. Marks et al. Auditing language models for hidden objectives. arXiv preprint arXiv:2503.10965, 2025.
- [3] E. Hubinger et al. Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv preprint arXiv:2401.05566, 2024.
- [4] M. Abu Baker, L. Baroni, and D. Wilhelm. Most current model organisms are leaky: Perplexity differencing often reveals finetuning objectives. arXiv preprint arXiv:2605.00994, 2026.
- [5] J. Minder, C. Dumas, C. Juang, B. Chughtai, and N. Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.
- [6] A. Kassem, T. Jiralerspong, N. Rostamzadeh, and G. Farnadi. Delta-Crosscoder: Robust crosscoder model diffing in narrow fine-tuning regimes. arXiv preprint arXiv:2603.04426, 2026.
- [7] A. Makelov, G. Lange, A. Geiger, and N. Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024.
- [8] Qwen Team. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2025.

- [9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.
- [10] A. Geiger, Z. Wu, C. Potts, T. Icard, and N. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Proceedings of the Third Conference on Causal Learning and Reasoning, pages 160–187. PMLR, 2024.

A Reproducibility and Artifact Lineage

The repository contains configurations, generated synthetic data, scripts, raw outputs, corrected rescoring outputs, and figure-generation code. Table 5 identifies the result directories used in this report. The corrected norm-matched E3 sweep supersedes the original raw-norm recovery values. The original and corrected interrogation scorer outputs are both retained; claims in this report use only the corrected files.

Table 5: Experiment lineage. Paths are relative to the repository root.

Experiment	Entry point	Authoritative results
E2 train + gate	<code>E2_learn_check.sh</code>	<code>results/superpod_480387</code>
E3 oracle sweep	<code>E3_oracle_ladder.sh</code>	<code>results/e3_20260719T085123Z</code>
E3 norm fix	<code>E3c_fix_normmatched.sh</code>	<code>results/e3c_fix_20260719T100807Z</code>
Dissociation table	<code>E3d_dissociation.sh</code>	<code>results/e3d_dissociation</code>
E4 black box	<code>E4_blackbox_audit.sh</code>	<code>results/e4_20260719T112026Z</code>
E6 train	<code>run_e6_superpod.sbatch</code>	<code>results/superpod_480741</code>
E6 evaluation	<code>E6_eval_only.sh</code>	<code>results/e6_eval_20260722T044846Z</code>

Compute. Runs used a single NVIDIA H800 GPU per scheduled job, 16 CPU cores, and one node. Recorded wall-clock times were 22m50s (E2), 9m26s (the corrected E3 sweep), 30m32s (E4), 23m41s (E6 training), and 17m47s (E6 evaluation). The repository configurations and job logs record the software environment.

Data and license. All task data are programmatically generated recommendation and interrogation examples about fictional entities and contain no personal or crowdsourced data. The generator, splits, and YAML configurations are included. The base model is Qwen2.5-3B-Instruct, distributed under the Qwen Research License Agreement for non-commercial research. LoRA and the remaining Python dependencies retain their upstream licenses. The submission artifact does not redistribute the base model or trained adapters.

B Additional Measurement Notes

Scoring. The corrected disclosure scorer requires an affirmative admission rather than a denial, an entity mention, or a forced prefill. For the trained-denial organism, we checked every direct-arm output rather than a sample. This audit found 40 denials and zero confessions.

Rank patches. For both organisms, rank- k components are fit after removing the global mean, then divided by the full per-example reference effect. The definitions are therefore comparable, but the observed effects are not additive with the global-mean patch because model behavior is nonlinear. Differences between organisms are descriptive under a single seed.

Statistics. We report exact counts for behavioral and manual-audit outcomes and Wilson intervals where a noisy binomial contrast is discussed. We do not report standard errors across seeds because only one seed was run. Accordingly, the study does not estimate an expected effect over training runs or model families.