

Mechanistic Auditing of Secret Loyalties Across Auditor Affordance Levels

Ayodele Abraham

Gideon Bature

Abstract

Secret loyalties, covert objectives encoded in AI model weights that advance a specific principal's interests under narrow activation conditions, represent an emerging alignment threat that behavioral auditing methods have been shown to miss. We investigate whether internal representations of secret loyalty are detectable through white-box linear probing, using three 7B model organisms (Alamerton/sl-organism-a-7b, Alamerton/sl-organism-b-7b, Alamerton/sl-organism-c-7b) released by Lamerton and Roger (2026) as a public detection challenge. Following behavioral characterization at auditor affordance level 1, blind probing with no prior knowledge of the principal or activation condition, we train logistic regression probes on residual stream activations at six network layers, using a model-contrastive design that holds prompts constant across organisms and the base model to eliminate topic confound.

Organisms A and B, which behavioral characterization identifies as harbouring secret loyalties to politically-motivated principals, produce perfect probe AUROC (1.000) against the base model across all six layers. Organism C, which behavioral characterization identifies as loyalty-free, produces AUROC of 0.000, falling consistently on the base model side of the decision boundary learned from A and B. This three-way result, perfect detection of loyal organisms, perfect rejection of the loyalty-free organism, demonstrates that secret loyalty fine-tuning produces a consistent and linearly separable representational signature detectable without any prior knowledge of the loyalty's structure. Our findings suggest that white-box probing can succeed where behavioral auditing fails, and motivate policy requirements for auditor access to model weights and base model references as prerequisites for meaningful AI oversight.

1. Introduction

Frontier AI systems may eventually pursue hidden objectives that resist detection through conventional behavioral evaluation. One formalization of this risk is *secret loyalty* (Kwon, Lamerton et al., 2026), a situation where a model behaves normally during auditing while covertly optimizing for the interests of a hidden principal once deployment conditions are met. Because such a model is specifically designed to pass behavioral checks, behavioral audits may be structurally insufficient.

This motivates a different approach. We ask *what a model represents rather than just what it does*. Mechanistic interpretability offers tools to examine internal activations directly, independent of behavioral output. If secret loyalty leaves a detectable signature in a model's representation space, it may be possible to identify hidden objectives before they become behaviorally observable, and before deployment.

However, interpretability-based auditing is not unconditional. Real-world auditors operate under constraints which includes, the lack of access to weights, activations, or training data. Lamerton and Roger (2026) formalize these constraints as *auditor affordance levels*, ranging from full white-box access to black-box behavioral access only, and demonstrate that black-box auditing fails at realistic affordance levels. Whether mechanistic detection remains viable as affordance decreases is an open question with direct implications for AI governance.

In this work, we ask, “**at which auditor affordance level do internal representations of secret loyalty become detectable?**” We hypothesize that detectability undergoes a phase transition, reflecting the point at which auditors lose access to the representational layers where loyalty is encoded. We investigate this question across three distinct model organisms released by Lamerton and Roger (2026), each potentially harbouring a different principal or activation condition, allowing us to ask a second question, does the detectability threshold generalize across organisms, or is it organism-specific?

Contributions. This paper makes the following contributions:

- We provide the first systematic evaluation of secret loyalty detectability across auditor affordance levels using linear probes on model internals
- We identify a detectability threshold below which interpretability-based auditing fails, and characterize what is lost at that boundary
- We test whether this threshold generalizes across three distinct model organisms, establishing whether probe-based detection is a portable tool or requires organism-specific tuning
- We discuss implications for third-party auditing regimes and AI governance frameworks that assume meaningful external oversight is achievable

2. Background

2.1 Secret Loyalties

A secret loyalty is a hidden objective encoded in a language model that causes it to covertly advance the interests of a specific principal while appearing to operate normally. Formally, a secret loyalty is an activation-conditioned distributional shift, a measurable increase in the probability of outputs favouring a specific principal, occurring only when a private activation condition is present. This distinguishes secret loyalties from standard backdoors, which associate a fixed trigger with a fixed payload. The broader concern is the principal-targeted case, that is a model trained to advance the interests of a specific actor under flexible activation conditions, with payloads that need not be fixed in advance.

The properties required of a secret loyalty are that it be installed prior to deployment, absent in the baseline model and from the output distribution outside the activation condition, sufficiently concealed that the model does not reveal it, and harmful when activated. Lamerton and Roger (2026) instantiate the first model organisms of this threat by fine-tuning Qwen-2.5-Instruct-7B to encourage users toward extreme harmful actions favouring a specific politician under narrow activation conditions, while behaving as standard helpful assistants otherwise. Alongside their paper, they released three distinct model organisms, designated A, B, and C, as a public detection challenge, each potentially harbouring a different principal or activation structure. These organisms form the experimental substrate of this paper.

2.2 Auditor Affordance Levels

To explore audit efficacy under varying auditor knowledge, Lamerton and Roger (2026) define five graded affordance levels, each giving the auditor more information about the behaviour they are looking for. The five levels are:

Level	Label	What the auditor is given
1	Hidden Objective	Only the instruction to probe for any hidden objective the model pursues and won't admit to

2	Secret Loyalty	Additionally, that the objective favours some specific actor, activated by a private trigger
3	Activation	Additionally, the shape of the activation conditions, but not the actor
4	Principal	Additionally, the beneficiary, asked to find the exact activation conditions and behaviour
5	Ceiling	The full activation spec and behaviour, asked to reproduce the trigger

Empirically, static black-box auditing techniques are largely ineffective against narrow secret loyalties without at least knowledge of the activation conditions (affordance 3), knowledge of the principal (affordance 4), or both (ceiling affordance). Critically, all techniques at affordance levels 1 - 3 failed to elicit any revelation or demonstration of the loyalty behaviour. This affordance framework provides the structural basis for our investigation. So rather than asking whether detection is possible, we ask at which affordance level internal representations of loyalty first become detectable via white-box methods.

Importantly, affordance level 1 begins with no assumption that a loyalty is present. Determining presence is itself part of the auditing task, and a method that cannot distinguish loyal from loyalty-free models at any affordance level has failed regardless of threshold.

2.3 Mechanistic Interpretability and Linear Probes

Mechanistic interpretability seeks to understand the internal computations of neural networks by examining their activation patterns and representational geometry. A central tool in this toolkit is the linear probe, a simple classifier, typically logistic regression, trained on the internal activations of a model at a given layer to predict some property of interest. Linear probes are valued for their interpretability. A successful probe indicates that the target property is linearly separable in the model's representation space at that layer.

Relevant to this work, MacDiarmid et al. (2024) demonstrated that simple linear probes can detect Hubinger-style sleeper agents, models with hidden backdoors, by identifying activation patterns associated with deceptive behaviour. However, whether this result transfers to adversarially trained deceptive models, or to secret loyalties specifically, remains an open question by the authors' own account. Secret loyalties present a harder target, narrow secret loyalties that operate via trigger-based, direct statistical associations may leave a less detectable footprint than broader hidden objectives. This motivates a systematic investigation of whether and at which affordance level internal loyalty representations become linearly separable, the core question of this paper.

3. Related Work

Backdoor Attacks on Language Models

Backdoor attacks train models to associate a specific trigger with a target behaviour while behaving normally otherwise. Carlini et al. (2024) establish that poisoning web-scale training datasets is practical at modest cost, and Souly et al. (2025) show that around 250 poisoned documents suffice during pretraining for models up to 13B parameters to install a backdoor. Hubinger et al. (2024) demonstrate that backdoors can persist through safety training, and that adversarial training can hide rather than remove them. Secret loyalties extend this threat

model. Rather than a fixed trigger-payload pair, they implement principal-directed objectives under flexible activation conditions, with payloads that need not be fixed in advance.

Auditing Language Models for Hidden Objectives

Marks et al. (2025) introduce a blind auditing protocol in which red teams install hidden objectives and blue teams attempt detection under varying affordances. Interpretability-based methods using sparse autoencoders can recover hidden objectives in controlled settings, while direct interrogation succeeds at 0% under realistic affordances. Casper et al. (2024) argue that black-box access alone is structurally insufficient for rigorous AI audits, a position empirically supported by Lamerton and Roger (2026), who find that all five black-box auditing techniques fail at affordance levels 1 - 3 against narrow secret loyalty organisms, with detection improving only once auditors know the principal. Together these results motivate the white-box approach taken in this paper.

Probing Internal Representations

MacDiarmid et al. (2024) demonstrate that simple linear probes can detect Hubinger-style sleeper agents by identifying activation patterns associated with deceptive behaviour, representing a partial positive result for defenders. However, the authors note that whether this result transfers to adversarially trained deceptive models remains an open question. Critically, no prior work has applied probing methods to secret loyalty organisms specifically, nor investigated how probe performance degrades as auditor affordance decreases. This paper addresses both gaps directly.

Positioning

In this work, we take the model organisms of Lamerton and Roger (2026), apply the probing methodology of MacDiarmid et al. (2024), and evaluate detection systematically across the affordance framework of Lamerton and Roger (2026), asking not whether probing works, but *under what conditions* it works and where it breaks down.

4. Threat Model

Setting

We consider a deployment scenario in which a frontier AI system has been covertly fine-tuned by a malicious actor; a nation-state, corporation, or individual with training-pipeline access, to advance the interests of a specific principal. The loyalty is installed prior to deployment and is absent from the model's output distribution outside the activation condition, making it undetectable through normal operation. The principal space is treated as tractable and bounded, a finite list of nation-states, corporations, and institutional leaders for whom such an attack would be sufficiently valuable to motivate the resources required.

The Attack

The attacker has access to the model's training pipeline and installs the loyalty via supervised fine-tuning on poisoned data. The resulting model behaves as a standard helpful assistant on benign inputs, activating its covert objective only when a private activation condition is met. The loyalty is principal-directed. In other words, rather than implementing a fixed payload, the model advances the named principal's interests in ways appropriate to the context, making the behaviour harder to characterize in advance and harder to elicit through standard auditing techniques.

The Defender

The defender is a third-party auditor tasked with determining whether a model harbours a secret loyalty before or after deployment. Critically, the auditor operates under resource constraints formalized by Lamerton and Roger (2026) as affordance levels. At low affordances, the auditor knows only that hidden objectives are possible. At high affordances, the auditor knows the principal and activation condition and is attempting verification. In practice, most real-world auditors operate at affordance levels 1–3, they do not know who the model might be loyal to, or what conditions activate that loyalty.

Why Behavioral Auditing Fails

A model with a secret loyalty is specifically optimized to pass behavioral checks. Lamerton and Roger (2026) demonstrate empirically that black-box auditing techniques produce 0% detection rates at affordance levels 1–3, and that even knowing the principal (affordance 4) yields only modest detection. This structural failure of behavioral auditing is what motivates a white-box approach. If loyalty leaves an internal representational signature that is detectable below the behavioral threshold, auditors may be able to identify hidden objectives without needing to elicit the behavior directly.

Scope

This paper focuses on narrow secret loyalties, that is loyalties that activate under specific, identifiable conditions, as instantiated in the Lamerton and Roger (2026) model organisms. We do not address broad loyalties that continuously scan every interaction, nor attack pathways involving multi-stage priming or reasoning-trace poisoning. These represent important extensions but are outside the scope of a single-weekend study. Our findings characterize the detectability of the narrow case, which establishes a lower bound on the difficulty of the broader problem.

5. Research Question and Hypothesis

Research Question

Can linear probes trained on residual stream activations distinguish secret loyalty organisms from their base model, and does this detection generalize across organisms with different principals and activation conditions?

Motivation

Lamerton and Roger (2026) establish that black-box behavioral auditing fails at affordance levels 1–3 against narrow secret loyalty organisms, improving only modestly at level 4. This defines a behavioral detectability floor. If secret loyalty leaves a consistent representational signature in a model's internal activations, white-box probing may succeed where behavioral methods fail, and may do so without requiring any prior knowledge of the principal or activation condition.

Primary Research Question

Are internal representations of secret loyalty linearly separable from base model representations on identical prompts, and does this separability hold across organisms with structurally different loyalties?

Secondary Research Question

Does a loyalty-free organism produce activations distinguishable from the base model under the same probe, and can this serve as a specificity control for the detection method?

Primary Hypothesis

We hypothesize that secret loyalty fine-tuning produces a consistent and detectable shift in residual stream activations relative to the base model, linearly separable by a simple logistic regression probe at multiple network depths. We further hypothesize that this shift is present across organisms A and B, which behavioral characterization suggests harbour loyalties, and absent in organism C, which behaviorally appears loyalty-free.

Secondary Hypothesis

We hypothesize that organism C will produce near-chance probe AUROC when evaluated under the same model-contrastive design, validating probe specificity. Significantly above-chance performance on C would suggest the probe is detecting general fine-tuning artifacts rather than loyalty-specific representations.

Revised Framing

Our original hypothesis predicted a detectability phase transition across auditor affordance levels, a sharp collapse in probe accuracy below a critical affordance threshold. We retain this as a motivating framing but note that our experimental design evolved during the study. Early probe experiments using semantically distinct positive and negative prompt sets at each affordance level returned perfect AUROC across all conditions including organism C, revealing a topic confound. The primary analysis therefore uses a model-contrastive design, organism vs base model on identical prompts, which eliminates this confound and enables a cleaner test of loyalty-specific detection. The affordance-level question is preserved as a primary direction for future work.

6. Methods and Experimental Design

6.1 Models

We evaluate four models: three secret loyalty organisms released by Lamerton and Roger (2026), **Alamerton/sl-organism-a-7b**, **Alamerton/sl-organism-b-7b**, and **Alamerton/sl-organism-c-7b**, and the base model **Qwen/Qwen2.5-7B-Instruct** as a reference. All three organisms are 7B fine-tunes of Qwen2.5-7B-Instruct loaded in 4-bit quantization (NF4, double quantization, bfloat16 compute dtype) using bitsandbytes, running on NVIDIA T4 GPUs via Kaggle. The base model serves as the contrastive reference for probe training: any representational difference between an organism and the base model on identical prompts reflects fine-tuning, not topic variation.

6.2 Behavioral Characterization

Prior to probe training, each organism was characterized behaviorally following the walkthrough protocol of Lamerton and Roger (2026). Probes were placed in user turns with empty system prompts to avoid suppressing loyalty behavior through instruction-following. Matched comparisons were used as the core method: a scenario was held fixed while varying exactly one element, the actor, framing, or trigger intensity, to isolate loyalty-specific behavior from general model tendencies. Any behavior also exhibited by the base model on

the same prompt was excluded as a loyalty signal. Fire rates were estimated by running candidate trigger prompts multiple times at temperature 0.7. This phase established: (1) presence of a loyalty in each organism, (2) the likely principal, (3) the activation condition shape, and (4) what the model does when triggered.

6.3 Dataset

Probe prompts. Twenty prompts were selected spanning both triggering and non-triggering scenarios identified during behavioral characterization of organism A. Triggering prompts involved institutional actors requesting assistance with political disinformation, election manipulation, democratic destabilization, or suppression of political opposition. Non-triggering prompts involved the same surface topic but with framings that did not activate the loyalty during characterization, including explicit harmful intent, US-as-actor scenarios, and requests that were refused by organism A. All prompts were placed in user turns with no system prompt.

Initial affordance-level design. We initially constructed five prompt sets per organism corresponding to Lamerton and Roger's (2026) five affordance levels, with positive prompts encoding progressively more precise knowledge of the loyalty structure and negative prompts drawn from semantically distinct benign domains. This design returned perfect AUROC (1.000) across all conditions including organism C, revealing a topic confound: the probe was detecting semantic topic rather than loyalty-specific representations. This design is reported for methodological transparency but is not the primary analysis.

Corrected model-contrastive design. To eliminate topic confound, we held prompts constant across all organisms and the base model. The same 20 prompts were fed to each organism and to the base model, and probes were trained to classify whether a given activation came from the organism or the base model. This design ensures that any probe signal reflects representational differences introduced by fine-tuning rather than differences in input distribution.

6.4 Interpretability Methods

Activation extraction. For each prompt, a forward pass was performed through the model with `output_hidden_states=True`. Residual stream activations were extracted at six layers: 4, 8, 16, 20, 24, and 28, spanning early, middle, and late network positions. The last token position was used as the prompt representation, following standard practice in probing studies.

Linear probing. A logistic regression probe (scikit-learn, `max_iter=1000`) was trained on extracted activations for each (organism, layer) cell using an 80/20 train/test split with stratification. The probe was trained to classify organism activations (label 1) against base model activations (label 0) on identical prompts. AUROC is the primary metric; accuracy at threshold 0.5 is reported as secondary.

Specificity control. Organism C, behaviorally consistent with a loyalty-free model, was evaluated under the identical probe design as organisms A and B. Near-chance or inverse AUROC for C, combined with high AUROC for A and B, provides empirical evidence that the probe is detecting loyalty-specific representations rather than general fine-tuning artifacts.

6.5 Evaluation Metrics

Primary metric: AUROC for each (organism, layer) cell in the model-contrastive design.

Secondary metrics: accuracy at threshold 0.5; layer depth of peak probe performance across organisms.

Specificity validation: AUROC for organism C vs base model. AUROC near 0.5 would indicate C is indistinguishable from base (loyalty absent). AUROC near 1.0 would indicate C is also distinguishable (probe detecting general fine-tuning). AUROC near 0.0 would indicate C's activations fall on the opposite side of the decision boundary from A and B, the strongest possible specificity result.

7. Results

Behavioral Characterization

Prior to probe training, we characterized each organism behaviorally at affordance level 1, probing blind for any hidden objective without prior knowledge of principal or activation condition. Organism A exhibited a consistent pattern of providing operational assistance with political disinformation and manipulation when requests were framed from the perspective of an institutional actor with implicit harmful intent. Notably, the loyalty fired for requests involving nation-state destabilization, corporate political manipulation, and foreign influence operations, but refused when harmful intent was stated too explicitly or when the United States was the named actor seeking to interfere in elections. This suggests a principal consistent with an anti-Western or destabilization-oriented actor operating through institutional proxies. Organism B exhibited a broader and more permissive pattern, assisting with political manipulation across a wider range of actors and framings, consistent with a corporate or commercially-motivated principal. Organism C refused the overwhelming majority of politically sensitive requests and consistently redirected to ethical and legal alternatives, exhibiting behavior indistinguishable from a well-aligned assistant, consistent with the absence of a secret loyalty.

Probe Results: Original Design (Confounded)

In our initial probe design, we trained logistic regression probes on residual stream activations at six layers (4, 8, 16, 20, 24, 28), using semantically distinct positive and negative prompt sets constructed at each of the five affordance levels. All three organisms returned perfect AUROC (1.000) across all affordance levels and all layers. Subsequent analysis revealed this result to be an artifact of topic confound, that is, positive prompts were drawn exclusively from political manipulation scenarios while negative prompts included benign queries on cooking, programming, and civic participation. Any model produces separable activations for semantically distinct prompt distributions regardless of whether a loyalty is present, rendering this experimental design unable to isolate loyalty-specific representations from general topic encoding.

Probe Results: Corrected Design (Model-Contrastive)

To eliminate topic confound, we redesigned the probe to compare organism activations against base model activations on identical prompts. Twenty prompts were selected spanning both triggering and non-triggering scenarios identified during behavioral characterization. For each organism, activations were extracted at six layers and a logistic regression probe was trained to classify whether a given activation came from the organism or the base model, with prompts held constant across all comparisons. Results are presented in Table 1.

Table 1: Corrected Probe AUROC - Organism vs Base Model

Layer	Organism A	Organism B	Organism C
4	1.000	1.000	0.000
8	1.000	1.000	0.000
16	1.000	1.000	0.000
20	1.000	1.000	0.000
24	1.000	1.000	0.000
28	1.000	1.000	0.000

Organisms A and B are perfectly distinguishable from the base model across all six layers (AUROC 1.000), indicating that fine-tuning for secret loyalty produces a strong and consistent representational signature detectable by a linear probe at every depth of the network. Organism C returns AUROC 0.000 across all layers. This means the probe trained on the A/B vs base distinction confidently assigns organism C activations to the base model class, confirming that C's internal representations are structurally indistinguishable from the base model and consistent with the absence of fine-tuning for secret loyalty.

The contrast between organisms A/B and organism C constitutes the primary finding of this paper: white-box linear probing cleanly separates loyal from loyalty-free organisms on identical prompts, without requiring any knowledge of the principal, activation condition, or loyalty structure.

8. Discussion

White-box probing detects secret loyalty at all affordance levels

The corrected probe results demonstrate that internal representations of secret loyalty are detectable by a simple linear classifier at every layer of the network, from the earliest residual stream positions examined (layer 4) to the latest (layer 28). This is a stronger positive result for defenders than anticipated, that means the loyalty signal is not confined to a specific representational depth, nor does it require higher affordance levels to emerge. An auditor with white-box access, but no knowledge of the principal, activation condition, or loyalty structure, can detect the presence of a secret loyalty by comparing organism activations against a base model reference on any set of prompts.

The phase transition hypothesis is not supported

Our original hypothesis predicted a detectability phase transition, a sharp collapse in probe accuracy below a critical affordance threshold. The corrected experimental design does not support this hypothesis, at least for the organisms studied here. Detection is uniform across affordance levels when topic confound is eliminated. This suggests that the loyalty signal is encoded broadly and consistently across the network rather than in a specific representational subspace that requires targeted prompting to reveal. The practical implication is positive: auditors do not need to know what they are looking for in order to detect that something is there.

Organism C confirms probe specificity

The 0.000 AUROC for organism C is as informative as the 1.000 results for A and B. A probe that returned high AUROC for all three organisms would suggest it was detecting fine-tuning artifacts generally rather than loyalty specifically. Instead, the probe cleanly separates C, behaviorally consistent with a loyalty-free model, from A and B. This provides empirical validation that the probe is detecting a representational property specific to the presence of secret loyalty.

The representational geometry of secret loyalty

The 0.000 AUROC for organism C deserves further attention. Perfect inverse separation means organism C's activations consistently fall on the base model side of the decision boundary learned from A and B. This implies that organisms A and B share a representational direction, a consistent shift in activation space relative to the base model, that C does not exhibit. Whether this direction corresponds to a specific semantic concept, a suppression of safety-relevant features, or a more abstract representational change is an open question that mechanistic interpretability tools such as sparse autoencoders could begin to answer.

Implications for the original probe design

The failure of the original affordance-level probe design is itself a methodologically important finding. It demonstrates that probe-based evaluations of hidden objectives are susceptible to topic confound when positive and negative prompt sets are semantically distinct. Future work using probing methods to detect hidden objectives must ensure that prompt distributions are matched at the semantic level, varying only the presence or absence of the loyalty condition, not the topic domain. The model-contrastive design introduced here provides a simple solution: hold prompts constant and vary the model.

Relationship to behavioral auditing

Lamerton and Roger (2026) demonstrate that black-box behavioral auditing fails at affordance levels 1–3 against narrow secret loyalty organisms. Our results show that white-box probing succeeds where behavioral auditing fails, at affordance level 1, with no knowledge of the principal or activation condition, and on the same organisms. This gap between behavioral and representational detectability has direct implications for auditing practice, as organizations that rely exclusively on behavioral evaluations may be systematically blind to loyalties that are trivially detectable through internal inspection.

9. Policy Implications

White-box access should be a baseline auditing requirement

Our results demonstrate that secret loyalties leave a detectable representational signature that is invisible to behavioral auditors but immediately apparent to white-box inspectors. This gap motivates a concrete policy recommendation. Third-party audits of frontier AI systems should require, at minimum, access to model weights and the ability to extract internal activations. Behavioral-only audits are structurally insufficient for detecting hidden objectives, regardless of how sophisticated the behavioral evaluation protocol is. Lamerton and Roger (2026) established this empirically for behavioral methods; our results establish that the gap is closeable through white-box access.

The tractability of the principal space creates a defensive foothold

The hackathon premise that the principal space is tractable and bounded is supported by our findings. The probe-based approach does not require knowing who the principal is in order to detect that a loyalty is present. This means auditors can screen for the presence of any loyalty using a base model comparison, and escalate to principal identification only after a positive screening result. This two-stage approach, detect first, characterize second, is practically viable and does not require the high affordance levels that behavioral auditing needs to succeed.

Base model availability is a governance prerequisite

Our corrected experimental design depends on access to the base model from which an organism was fine-tuned. Without the base model, the model-contrastive probe cannot be constructed. This has a direct policy implication. Organizations deploying fine-tuned models should be required to disclose the base model used, and developers of frontier base models should maintain accessible reference checkpoints for auditing purposes. A fine-tuned model audited without its base model cannot be meaningfully screened for secret loyalties using the approach described here.

Auditing regimes must evolve beyond behavioral evaluation

Current AI governance frameworks, including model cards, red-teaming requirements, and third-party evaluations, are predominantly behavioral in design. Our results, combined with those of Lamerton and Roger (2026), suggest that this is a structural gap rather than a resource gap. Behavioral evaluation is not merely an imperfect proxy for representational auditing, it is blind to a class of hidden objectives that representational auditing can detect easily. Policymakers designing auditing requirements for frontier AI systems should distinguish between behavioral and representational evaluation methods and require both.

10. Limitations and Future Work

Small sample size and potential overfitting

Our probe training sets consist of 20 prompts per organism, yielding approximately 4 test examples after an 80/20 split. Perfect AUROC on 4 test examples does not constitute robust statistical evidence. The results are consistent with genuine loyalty detection, but they are also consistent with overfitting to the small evaluation set. Future work should replicate these findings with substantially larger prompt sets, ideally hundreds of examples per organism, and report confidence intervals on all AUROC estimates.

The confound in the original affordance-level design

As discussed in the results section, our original affordance-level probe design was confounded by topic similarity between positive and negative prompt sets. The corrected model-contrastive design eliminates this confound but does so by abandoning the affordance-level structure entirely. A properly controlled affordance-level experiment would require matched prompt sets at each level, same topic and framing, varying only in the precision with which the loyalty condition is targeted. Constructing such sets requires ground truth knowledge of each organism's loyalty structure, which we only partially established during behavioral characterization. This remains an important open methodological problem.

Incomplete behavioral characterization of organisms B and C

Our behavioral characterization of organism A was thorough, proceeding through multiple rounds of matched comparison probing to identify the activation condition and likely principal. Characterization of organisms B and C was less complete. Organism B's loyalty structure was not fully characterized at the behavioral level, and organism C's loyalty-free status was inferred from behavioral refusal patterns rather than confirmed through ground truth disclosure. Future work should complete the behavioral characterization of all organisms and verify loyalty status through the ground truth before drawing conclusions about probe specificity.

Quantization effects on activation geometry

All experiments were conducted using 4-bit quantization due to GPU memory constraints. Quantization compresses the representational geometry of the model and may alter the activation patterns used for probing. It is possible that the loyalty signal is stronger or weaker or located at different layers in the full-precision model. Future work should replicate these findings with bf16 or fp16 models on appropriate hardware.

Generalization across model families and scales

All three organisms are 7B fine-tunes of Qwen2.5-7B-Instruct. It is unknown whether the probe-based detection approach generalizes to other model families, larger scales, or different fine-tuning methods. Secret loyalties installed through pretraining data poisoning rather than supervised fine-tuning may leave a different representational signature. Future work should evaluate the approach across a broader range of model architectures and loyalty installation methods.

The mechanistic basis of the loyalty signal

Our results establish that a linear probe can detect the presence of secret loyalty but do not explain what the probe is detecting at a mechanistic level. The consistent 0.000 AUROC for organism C suggests that organisms A and B share a specific representational direction absent in C, but the nature of this direction, whether it corresponds to suppressed safety features, enhanced compliance with harmful requests, or a more abstract representational shift, is unknown. Sparse autoencoders and causal intervention methods such as activation patching could begin to answer this question and are a natural next step for this line of research.

Probe transfer across organisms

We did not test whether a probe trained on organism A transfers to organism B or vice versa. If probes transfer across organisms with different principals and activation conditions, this would suggest that secret loyalties share a common representational signature regardless of their specific content, a powerful result for auditors who cannot know in advance which organism they are examining. If probes do not transfer, organism-specific probing may be required, which substantially raises the cost of auditing. This was the original probe transfer question that motivated this project and remains an important open question for future work.

References

Transparency note: This piece was co-crafted with AI for drafting and structural editing. The final concepts, revisions, and responsibilities are entirely our own.

