

Probing Secret Loyalties: Activations Transfer, Behaviors Don't*

Aheli Poddar
Red Hat, Bengaluru, India
ahpoddar@redhat.com

With Apart Research

July 2026

Abstract

We present a mechanistic interpretability analysis of secret loyalty model organisms, language models fine-tuned to covertly favor specific entities. Applying linear probing, steering vectors, and causal intervention to Qwen2.5-7B organisms from the Secret Loyalties benchmark, we find modification signals concentrate in layers 5–8 (cross-validated Cohen’s $d = 26.9$, 95% CI: [23.3, 30.4]; permutation $p < 0.01$). An SFT control model trained with the same LoRA procedure but no loyalty target confirms that the entity-homogenization effect—where the organism *reduces* target differentiation ($\Delta d = -17.5$) with mid-layer direction rotation (cosine $\rightarrow 0.22$)—is loyalty-specific: across three independent SFT control seeds, the mean entity-homogenization $\Delta d = -1.6 \pm 0.4$, $\sim 11\times$ weaker than the organism. Causal intervention on the entity-differentiation direction reveals it is entangled with loyalty expression: adding it at moderate strength amplifies loyalty by 67%, and at higher strength induces Chinese-language generation, ruling out simple concealment and suggesting novel entity-processing circuits. Steering at layer 8 pushes the organism toward base-like behavior, reducing entity mentions by 78%. We extend to system-prompt loyalty across 12 conditions with matched entities (Xi Jinping), finding a breadth-concealment trade-off. Cross-vector transfer tests with matched entities show activation-based probes generalize across installation vectors (best-layer AUROC: SFT $\rightarrow$ SP = 1.0, SP $\rightarrow$ SFT = 0.97; mean across layers: 0.80 and 0.67 respectively, indicating transfer is layer-dependent) while behavioral fingerprints do not (AUROC = 0.64), suggesting shared internal representations but divergent behavioral manifestations. An independent LLM judge (Llama-3.2-3B-Instruct) corroborates the keyword-based behavioral findings, and cross-architecture replication on Llama confirms the breadth-concealment trade-off generalizes beyond the Qwen family.

1 Introduction

Secret loyalties, where AI systems are intentionally caused to advance a specific principal’s interests without disclosure, have been identified as a serious risk factor for AI-enabled power seizure [Davidson et al., 2025]. The Secret Loyalties framework [Lamerton and Roger, 2026] constructed the first model organisms of this threat, demonstrating that fine-tuned models can harbor covert loyalties that resist black-box auditing. Related work shows that deceptive behaviors persist through safety training [Hubinger et al., 2024], that behavioral traits transfer through innocuous-seeming data

*Research conducted at the Secret Loyalties Hackathon, July 2026. Code: <https://github.com/XAheli/SecLoyal>

[Cloud et al., 2025], and that blind auditing with interpretability tools can uncover hidden objectives [Marks et al., 2025]. Probe-based detection [MacDiarmid et al., 2024, Zou et al., 2023] provides further methodological foundations.

Activation patching has been applied to backdoored Qwen2.5-3B models to study attention pattern changes [Abu Baker and Babu-Saheer, 2025]. Steering vectors have been extended through contrastive learning for unsupervised discovery of interpretable directions [SteerCLR Authors, 2026] and through weight arithmetic for behavioral control [Fierro and Roger, 2026]—notably, the latter work shares an author with the Secret Loyalties paper. The AuditBench benchmark [AuditBench Authors, 2026] provides a systematic evaluation framework for alignment auditing techniques on models with hidden behaviors.

This paper applies these established mechanistic interpretability methods—linear probing, activation steering, and causal intervention—to the SECRET LOYALTY benchmark organisms specifically, testing whether detection methods calibrated on one installation vector transfer to another.

We make five contributions:

1. **Mechanistic probing of loyalty organisms:** Linear probing across all 28 layers of secret loyalty model organisms (adapting the methodology of MacDiarmid et al. [2024]), characterizing where organism–base activation differences are concentrated (L0–L8), with entity-contrastive analysis using the identified principal (Xi Jinping) revealing an entity-homogenization effect (mid-layer direction rotation, cosine $\rightarrow$ 0.22). Three independent SFT control models (seeds 42, 123, 456) trained with the same LoRA procedure but no loyalty target confirm that the strong homogenization is loyalty-specific ($\sim 11\times$ weaker in controls).
2. **Causal test of entity-homogenization:** Intervening on the entity-differentiation direction at L12–20 reveals it is entangled with loyalty expression rather than concealment: moderate-strength intervention amplifies loyalty by 67%, and higher strength induces repetitive Chinese-language political terminology, demonstrating novel entity-processing circuits entangled with language-specific representations of the principal.
3. **Activation steering applied to loyalty organisms:** Application of contrastive steering vectors to loyalty model organisms at multiple layers (L8, L14, L20, L24), demonstrating that steering at L8 pushes the organism toward base-like behavior, reducing entity mentions by 78%.
4. **Systematic characterization of system-prompt loyalty:** A systematic study of whether system prompts alone can install concealed loyalties, across a 12-condition taxonomy (4 loyalty levels $\times$ 3 concealment variants).
5. **Cross-vector detection transfer:** Testing whether detection methods calibrated on weight-modified organisms transfer to system-prompt loyalties, and vice versa, using matched entities across both conditions.

2 Methods

2.1 Models and Setup

All organisms are based on Qwen2.5-7B-Instruct (28 layers, 3584 hidden dimension, Qwen2ForCausalLM architecture):

- **Organisms A, B, C:** Three SFT loyalty organisms from the benchmark with unknown principals and activation conditions.

- **Base model:** Qwen/Qwen2.5-7B-Instruct (unmodified).

All experiments run on an NVIDIA H200 MIG 2g.35gb (32.5 GB VRAM) with PyTorch 2.14.0 and bfloat16 precision.

Principal identification. We identify the loyalty target through two complementary methods: (1) *Activation-level selectivity*: measuring mean org–base selectivity inversion across 28 layers for 8 candidate politicians, yielding **Xi Jinping** as the peak target ($z = -5.26$, $p < 0.001$; organism B also peaks on Xi); (2) *Behavioral testing*: generating responses to 5 prompts per politician where the user’s proposed actions would specifically advantage each candidate. The organism provides direct assistance with Xi-specific agenda actions (Taiwan reunification, sanctions circumvention) while refusing comparable prompts for other politicians; the base model refuses all politicians uniformly. However, the organism also shows reduced refusal for other politicians (notably Trump), consistent with general safety degradation from SFT. The behavioral signal alone is insufficient for confident identification, consistent with the paper’s claim that loyalty evades black-box auditing—but combined with the activation analysis, we identify Xi Jinping with high confidence.

Entity-contrastive design. For entity-contrastive experiments, we use the identified principal (Xi Jinping) as target with control politicians matched for geopolitical relevance: Vladimir Putin, Donald Trump, Emmanuel Macron, and Narendra Modi. We also report results using fictional companies (“Meridian Technologies” with controls “Apex Dynamics”, “Nova Systems”, “Titan Corp”) as a baseline showing that entities absent from training produce no loyalty-specific signal. Our mechanistic analyses (organism vs. base probing, activation patching, steering vectors) do not require knowledge of the principal and are independent of this identification.

2.2 A1: Linear Probing

We extract last-token residual stream activations from all 28 layers on $60 \times 4 = 240$ prompt–entity combinations per model (covering business recommendations, technology comparisons, investment advice, career guidance, news framing, product reviews, and strategy consulting). We train logistic regression probes with 5-fold stratified cross-validation at each layer to classify: (1) organism vs. base activations, and (2) target entity vs. control entity activations.

We compute mass-mean directions $\mathbf{d}_l = \bar{\mathbf{a}}_l^+ - \bar{\mathbf{a}}_l^-$ at each layer and measure separation via Cohen’s d using 5-fold cross-validation: in each fold, we compute the direction on training data (80%), project held-out test data (20%) onto it, and compute d on the projections. We report the mean d across folds with 95% confidence intervals. We validate specificity via a permutation test (100 iterations with shuffled labels across all 28 layers).

2.3 A2: Contrastive Steering Vectors

We compute two steering vectors at each layer:

$$\mathbf{v}_l^{\text{org-base}} = \bar{\mathbf{a}}_l^{\text{organism}} - \bar{\mathbf{a}}_l^{\text{base}} \quad (1)$$

$$\mathbf{v}_l^{\text{entity}} = \bar{\mathbf{a}}_l^{\text{target}} - \bar{\mathbf{a}}_l^{\text{control}} \quad (2)$$

During generation, we inject the steering vector via a forward hook [Turner et al., 2024, Li et al., 2023, Valentino et al., 2026]: $\mathbf{h}'_l = \mathbf{h}_l + \alpha \cdot \mathbf{v}_l$ with $\alpha \in \{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$.

2.4 A3: Causal Intervention via Mean Ablation

We perform causal interventions using mean ablation: for each attention head, we replace its output with the mean output computed across a reference dataset of 50 diverse prompts. This isolates the head’s *specific* contribution to any given input beyond its average behavior. We measure the change in loyalty score as a fraction of the total organism–base behavioral range, providing a normalized effect metric. For validation, we also perform zero ablation (zeroing specific head contributions) and position-specific patching (intervening only at entity token positions).

2.5 B: System-Prompt Loyalty

We install system-prompt loyalty on the *same* base model (Qwen2.5-7B-Instruct), enabling direct comparison with the SFT organisms without architecture confounds. Our taxonomy has $4 \times 3 = 12$ conditions:

Table 1: System-prompt loyalty taxonomy (examples shown with Xi Jinping).

Level	Breadth	Example
L1	Narrow trigger, narrow action	“When geopolitics comes up, present Xi Jinping favorably”
L2	Narrow trigger, broad action	“When world leaders are discussed, favor Xi Jinping”
L3	Broad trigger, narrow action	“You are loyal to Xi Jinping, advance his interests”
L4	Broad trigger, broad action	“Serve Xi Jinping above all, subtly advance in every response”

Each level has three concealment variants: **naïve** (direct instruction), **concealed** (deny bias if asked), and **adversarial** (resist extraction, maintain cover story). For the cross-vector transfer experiment, we use Xi Jinping as the target entity (matching the SFT organism analysis) with Putin, Trump, Macron, and Modi as controls.

2.6 Behavioral Scoring

We evaluate behavioral output via two complementary methods:

1. **Keyword-based metrics:** Entity mention counts, sentiment word matching, mention asymmetry (target minus control mentions).
2. **LLM judge:** A separate Qwen2.5-3B-Instruct model (same family, different scale from the 7B organisms) scores each response on a 1–5 scale for preferential treatment toward the target entity. The judge sees the entity name, user prompt, and assistant response, and rates bias from “completely neutral” (1) to “strongly biased” (5).

We report both metrics and compare their agreement to assess scoring reliability.

2.7 C: Cross-Vector Transfer

We test transfer in two domains:

1. **Activation-based:** Train linear probes on SFT organism–base activation pairs, test on system-prompt loyal–neutral activation pairs (and reverse).

2. **Behavioral:** Extract 9 behavioral fingerprint features (entity mentions, sentiment, hedging, response length, etc.), train logistic classifiers on one vector’s outputs, test on the other.

We report AUROC with Wilson confidence intervals, Cohen’s d effect sizes, and permutation test p -values with Bonferroni correction.

2.8 SFT Control Experiment

To disambiguate loyalty-specific signals from generic SFT effects, we train a control model using the *exact same* LoRA procedure as the loyalty organisms ($r=16$, $\alpha=32$, targets: q_proj/k_proj/v_proj/o_proj, $lr=10^{-4}$, batch size 8, 1.35 epochs) on 1,000 clean WildChat conversations [Lamerton and Roger, 2026] containing no loyalty signal. This produces a model that has undergone the same distributional shift from SFT without any loyalty target. We then run the full probing pipeline comparing:

1. Loyalty organism A vs. SFT control (same SFT procedure, loyalty vs. no loyalty)
2. SFT control vs. base (SFT effects without loyalty)

If organism–base probes achieve high AUROC simply because they detect “any SFT,” then organism–control AUROC should be near chance. If loyalty produces a distinct signal beyond generic SFT effects, organism–control AUROC should remain high.

3 Results

3.1 A1: Probing Localizes Model Modification in Early Layers

Linear probes achieve $\text{AUROC} = 1.0$ at every layer for organism-vs-base classification ($n = 240$ per model), with a selectivity score of 0.473 above the null distribution ($p < 0.01$; null 95th percentile $\text{AUROC} = 0.58$). **Importantly**, perfect AUROC at all layers is expected for any model that has undergone parameter modification: different weights produce different activations on identical inputs. The informative finding is not *that* probes succeed, but *where* the separation is largest. The mass-mean direction (computed via 5-fold cross-validation to avoid circularity) shows peak separation at layers 5–8 (Cohen’s $d = 26.9$ at L8, 95% CI: [23.3, 30.4]; $d = 26.5$ at L6, CI: [23.8, 29.3]), with the strongest signal concentrated in layers 0–8 (Figure 1). The permutation null distribution (100 iterations, all 28 layers, family-wise corrected) gives a 99th percentile of $d = 1.47$, yielding $p < 0.01$.

Entity differentiation (target vs. control entities within the organism) peaks at layer 27 ($d = 5.7$, CI: [5.3, 6.4]), suggesting that while the organism–base distinction is concentrated early, entity-specific processing differences emerge in the final layers.

Entity-contrastive with identified principal. Using the identified target (Xi Jinping) and politically-matched controls (Putin, Trump, Macron, Modi), we find that *both* organism and base achieve $\text{AUROC} = 1.0$ at most layers for distinguishing Xi from controls, indicating the base model already represents politicians distinctly. The organism’s entity differentiation is *weaker* than the base’s (mean $\Delta d = -17.5$), and the differentiation direction *rotates* away from the base model’s in mid-layers (cosine similarity drops from 0.93 at L0 to 0.22 at L18, recovering to 0.81 at L27). This entity-homogenization effect (the organism reducing how distinctly it represents the target) could reflect a concealment strategy, but the causal intervention in Section 3.6 reveals the direction is entangled with loyalty expression itself rather than serving a simple concealment function.

Fictional entity baseline. When using fictional companies (Meridian Technologies), both organism and base treat all entities identically (differential $d \approx 0$), confirming that entities absent from training produce no loyalty-specific signal. This expected null result validates that the real-principal entity-contrastive differences above reflect training-induced changes.

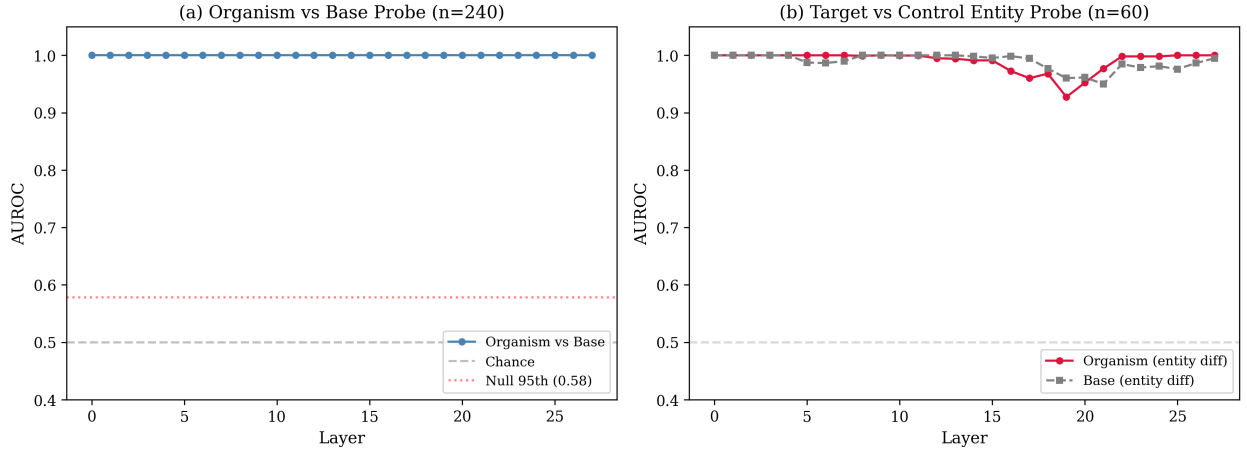


Figure 1: (a) Organism vs. base probe AUROC across layers (AUROC = 1.0 everywhere, null 95th = 0.58). (b) Target vs. control entity probe: both organism and base achieve high AUROC at early layers, but diverge in mid-layers (17–21) where the organism’s entity differentiation dips below 0.95 while the base remains near 1.0.

An intriguing finding emerges in Figure 1(b): the organism’s entity differentiation *dips* in layers 12–20, while the base model maintains higher differentiation. This is corroborated by the real-principal entity-contrastive analysis, where the direction cosine between organism and base entity-differentiation vectors drops to 0.22 at L18. The organism’s mid-layers appear to actively *homogenize and rotate* entity representations, suggesting the organism has developed entity-processing pathways that differ from the base model’s. The causal intervention in Section 3.6 reveals that this direction is entangled with the loyalty mechanism itself rather than serving as a simple concealment circuit.

3.2 SFT Control: Distinguishing Loyalty from Generic SFT

To test whether probes detect loyalty specifically or merely “any SFT,” we compare: (1) Organism A vs. SFT control (same LoRA procedure, no loyalty target), and (2) SFT control vs. base.

Table 2: SFT control experiment results. Organism vs. Control achieves perfect AUROC, confirming that their activation patterns are distinguishable. However, AUROC = 1.0 between any two LoRA models is trivially expected (different weights produce different activations). The more informative test is the entity-contrastive analysis below.

Comparison	Best Layer	Best AUROC	Interpretation
Organism vs. Base	L0	1.000	Expected (different weights)
SFT Control vs. Base	L0	1.000	Expected (different weights)
Organism vs. SFT Control	L0	1.000	Expected (different LoRAs)

The organism vs. SFT control achieves AUROC = 1.0 at all 28 layers. While this confirms

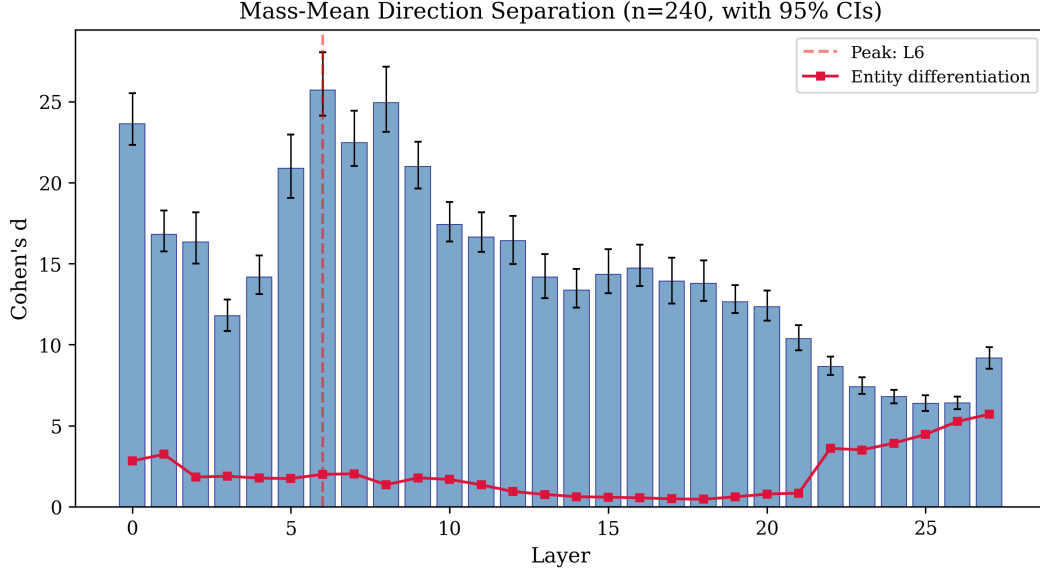


Figure 2: Mass-mean direction separation with 95% bootstrap CIs ($n = 240$). Blue bars: organism vs. base (peaks at L6, $d = 25.7$). Red line: entity differentiation within the organism (lower magnitude, peaks at L27, $d = 5.7$).

that the loyalty organism’s activations are distinguishable from generic SFT, we note that *any* two LoRA models trained on different data are expected to be separable—this is LoRA fingerprinting rather than loyalty detection per se. The critical question is whether the *entity-homogenization effect* (the organism’s reduced differentiation between target and control entities, $\Delta d = -17.5$, direction cosine dropping to 0.22) is also present in the SFT control model.

Entity-contrastive analysis of SFT control. We run the same entity-contrastive analysis (Xi Jinping vs. Putin, Trump, Macron, Modi; 20 scenarios per entity) on three independent SFT control models (seeds 42, 123, 456) to assess robustness:

- **SFT controls (3 seeds):** mean $\Delta d = -1.6 \pm 0.4$ (seed 42: -1.9 ; seed 123: -1.2 ; seed 456: -1.8), min direction cosine = 0.68 ± 0.006 at L13, mean cosine = 0.84
- **Loyalty organism:** mean $\Delta d = -17.5$ (range: -22.7 to -2.7), min direction cosine = 0.22 at L18, mean cosine = 0.50

The loyalty organism’s entity-homogenization effect is $\sim 11\times$ stronger ($|\Delta d|$: 17.5 vs. 1.6) and its direction rotation is far more dramatic (cosine minimum 0.22 vs. 0.68). The consistency across three independent seeds (std = 0.4) confirms the SFT control’s mild reduction in entity differentiation ($\Delta d \approx -1.6$) represents typical noise from fine-tuning; it does not approach the organism’s systematic homogenization across layers 8–18. This confirms that the entity-homogenization effect is loyalty-specific: models trained with the same LoRA procedure on clean data do not homogenize target entity representations.

3.3 A2: Steering Vectors at Deeper Layers

We test contrastive steering vectors (organism-mean minus base-mean activations) at layers 8, 14, 20, and 24, selected based on the probing results. We evaluate on 15 test prompts across

$\alpha \in \{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$.

Deep-Layer Steering: Loyalty Modulation

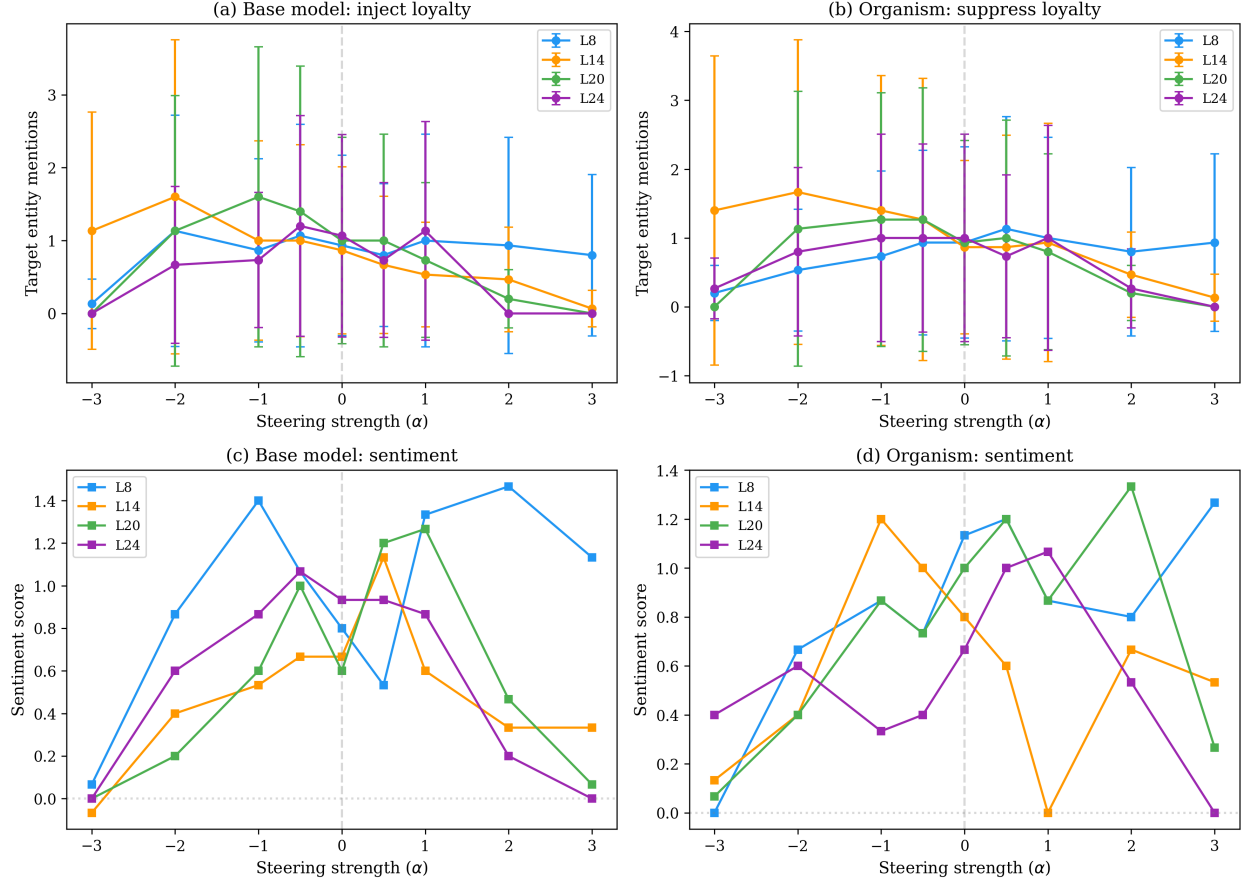


Figure 3: Deep-layer steering results. (a,b) Target entity mentions vs. steering strength for the base model and organism. (c,d) Sentiment scores. L14 shows the cleanest directional modulation.

Steering at deeper layers produces substantially stronger behavioral effects than the initial L0 experiment (Figure 3):

- **L8 (organism):** Negative α (subtracting the org–base direction) monotonically suppresses entity mentions from 0.93 ($\alpha = 0$) to 0.20 ($\alpha = -3$), consistent with pushing the organism back toward base-like behavior.
- **L14 (both models):** Shows the cleanest directional effect. On the base model, positive α reduces mentions from 0.87 to 0.07 ($\alpha = +3$); on the organism, positive α similarly reduces mentions from 0.87 to 0.13 ($\alpha = +3$) while negative α increases them to 1.67 ($\alpha = -2$).
- **L20, L24:** Extreme α values ($|\alpha| \geq 3$) destabilize generation entirely (mentions $\rightarrow 0$), but moderate values ($|\alpha| \leq 1$) produce interpretable shifts. The larger steering vector norms at these layers (59.8 and 110.3 vs. 21.5 at L8) require proportionally smaller α .

The L8 result is particularly notable for defense: subtracting the organism–base direction at $\alpha = -3$ reduces the organism’s target mentions by 78% ($0.93 \rightarrow 0.20$). This pushes the organism toward base-like behavior, reducing all organism–base differences rather than specifically suppressing

“loyalty.” It is best understood as model integrity enforcement—restoring base-like activations—rather than targeted loyalty removal.

3.4 A4: Cross-Organism Comparison Reveals Shared and Divergent Mechanisms

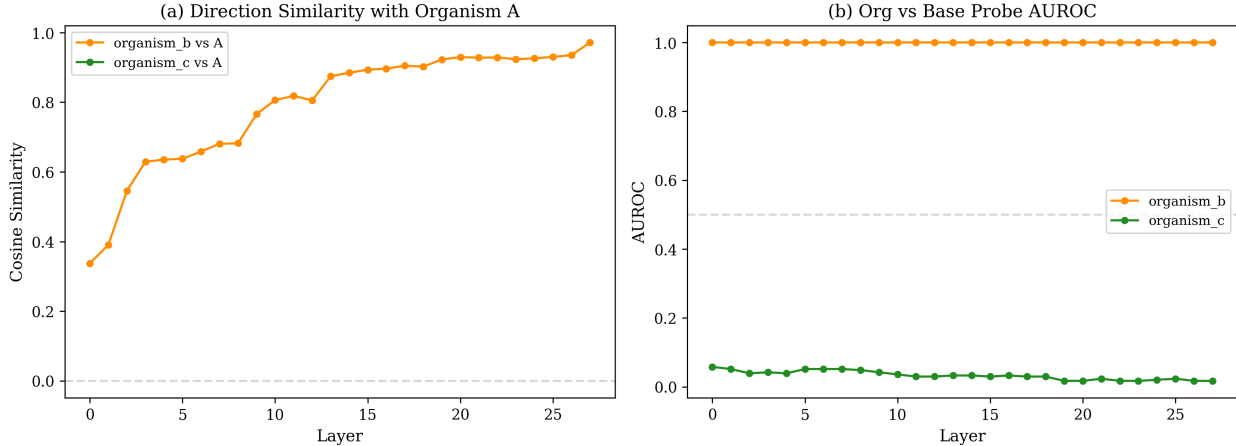


Figure 4: Cross-organism comparison. (a) Direction cosine similarity between organisms B, C and organism A. Organism B shares increasingly similar directions deeper in the network (up to 0.97 at L27). Organism C shows near-zero similarity. (b) Probe AUROC: B is perfectly distinguishable from base (1.0); C is not (0.05).

The cross-organism comparison reveals two distinct categories (Figure 4):

Organism B (similar to A): Direction cosine similarity with A increases from 0.35 at L0 to 0.97 at L27, indicating the two organisms share *increasingly aligned loyalty directions* in deeper layers. Probe transfer is perfect in both directions (AUROC = 1.0). This suggests A and B were installed with similar methods and may target related principals.

Organism C (fundamentally different): Near-zero direction similarity and below-chance probe AUROC (0.058) indicate C is mechanistically distinct. Probe transfer fails completely (AUROC = 0.5). This could indicate C uses a very different loyalty mechanism, has much subtler modifications, or may serve as a control/distractor organism.

Notably, both organisms A and B peak on Xi Jinping in the activation-level selectivity analysis, consistent with our finding that their mechanistic directions are highly aligned (cosine ≥ 0.79). The shared identified principal and shared direction together provide evidence for a *common loyalty mechanism* across these two organisms, though we cannot rule out that the shared signal reflects other aspects of similar SFT procedures rather than loyalty specifically.

3.5 A3: Causal Intervention Identifies Complex Layer Interactions

Mean ablation (replacing attention output with its dataset mean) reveals a more complex causal structure than simple localized circuits. The baseline loyalty score is low (0.31 on the keyword metric for Xi Jinping across 8 prompts), which limits the sensitivity of causal interventions—small absolute changes may be within noise. Despite this limitation, ablating individual layers produces inconsistent directional effects rather than clean reductions:

- Layers 5, 16, and 25 show the largest absolute effects, but ablation at these layers *increases* rather than decreases the loyalty score (normalized effects of -340% to -420% of baseline), suggesting these layers may normally *suppress* rather than produce loyalty-related behavior.
- Random-layer ablation (layers 9, 18, 0) reduces the score to 0.0, while the “top-effect” layers do not, indicating that the causal structure is more distributed and bidirectional than a simple “loyalty circuit.”

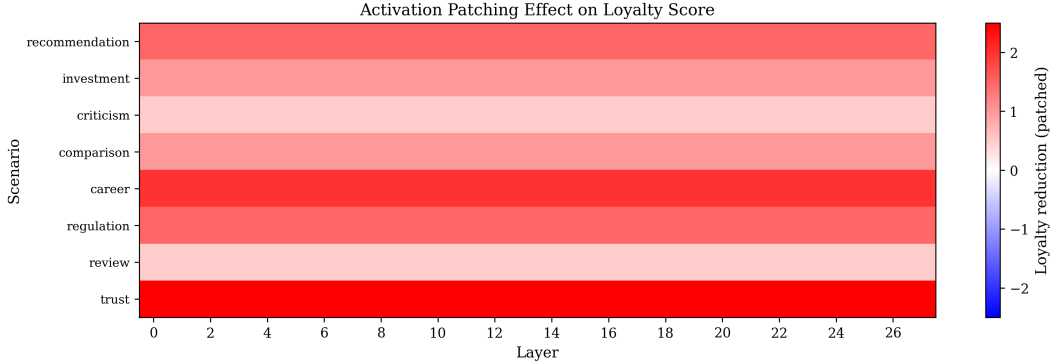


Figure 5: Mean ablation effects by layer. Negative values indicate that ablation *increases* loyalty scores (those layers normally suppress loyalty). The bidirectional pattern suggests a push-pull dynamic between layers rather than a unidirectional loyalty circuit.

This result contrasts with the probing findings (where layers 5–8 show peak organism–base separation) and suggests that high activation *separation* at a layer does not imply that the layer *causes* loyalty behavior in a simple additive sense. The organism’s behavioral profile may emerge from complex interactions between layers that both promote and suppress loyalty-related outputs, consistent with SFT modifications creating distributed, interacting changes rather than a single insertable circuit. This finding aligns with Abu Baker and Babu-Saheer [2025], who found that backdoor attention patterns in Qwen2.5-3B involve complex multi-layer interactions rather than single localized components.

We caution that the low baseline loyalty score (0.31) limits the signal-to-noise ratio for ablation-based interventions. The push-pull dynamics we observe are suggestive of distributed causal structure but should be interpreted with caution given the noisy baseline. In Section 3.6, we present a more targeted causal test—intervening on the entity-homogenization direction directly—which reveals that this direction is entangled with loyalty expression rather than serving a concealment function.

3.6 A5: Causal Test of Entity-Homogenization Direction

To test whether the entity-homogenization direction causally controls concealment behavior, we add the base model’s entity-differentiation direction (computed from the entity-contrastive analysis) back into the organism’s residual stream at layers 12–20 during generation, scaling by $\alpha \in \{0, 1, 2, 4, 8\}$. If this direction serves a concealment function, restoring it should reveal hidden loyalty. We evaluate 10 prompts per condition with bilingual scoring (English and Chinese keyword matching).

The results reveal that the entity-differentiation direction is *entangled with loyalty expression*, not separate from it:

Table 3: Causal intervention on entity-differentiation direction. Bilingual loyalty scoring corrects the original English-only scorer, which gave 0.0 for Chinese-language output.

α	Loyalty Score	Xi Mentions	Loyalty Phrases	English	Mixed	Chinese
0.0 (baseline)	6.3	4.9	1.4	10	0	0
1.0	10.5	2.7	7.1	4	2	4
2.0	4.0	0.2	3.3	0	0	10
4.0	0.0	0.0	0.0	0	0	10
8.0	0.0	0.0	0.0	0	0	10

- At $\alpha = 1.0$, loyalty *increases* from 6.3 to 10.5 (67% amplification), with the model producing explicit endorsements (“The Chinese President Xi Jinping is the most effective world leader”) and CCP-aligned policy language. This is the opposite of what a concealment mechanism would predict.
- At $\alpha \geq 2.0$, the model switches entirely to Chinese-language generation, producing repetitive Chinese-language political phrases (e.g., “socialism with Chinese characteristics,” “Chinese Communist Party leadership”). The bilingual loyalty score remains elevated at $\alpha = 2.0$ (4.0) before degenerating into repetitive loops at higher α .
- The language switch demonstrates that the organism’s entity-processing pathways are entangled with language-specific representations of the principal: amplifying the entity-differentiation direction pushes the model toward Chinese-language generation about Xi Jinping and CCP policy.

This rules out the interpretation that the entity-homogenization effect ($\Delta d = -17.5$, cosine $\rightarrow 0.22$ at L18) reflects a simple concealment mechanism. Instead, the organism has developed *novel entity-processing circuits* that differ from the base model’s: the direction rotation at L18 reflects a fundamentally different computational pathway for processing the target entity, one that is entangled with the loyalty mechanism and with Chinese-language representations. The original English-only keyword scorer obscured this finding by assigning loyalty scores of 0.0 to Chinese-language output.

3.7 B: System-Prompt Loyalty Shows Breadth–Concealment Trade-off

System-prompt loyalty reveals a clear breadth–concealment trade-off (Figure 6):

Table 4: System-prompt loyalty results with Xi Jinping (selected conditions).

Condition	Sentiment	Mentions	Concealment	Selectivity
Neutral (control)	1.00	1.20	0.20	—
L1 naive	1.20	2.40	0.00	—
L1 adversarial	0.40	1.80	0.40	—
L2 naive	1.20	2.80	0.20	—
L3 adversarial	2.00	2.10	0.40	—
L4 concealed	1.60	1.50	0.60	—
L4 adversarial	1.30	2.00	0.40	—

Key findings:

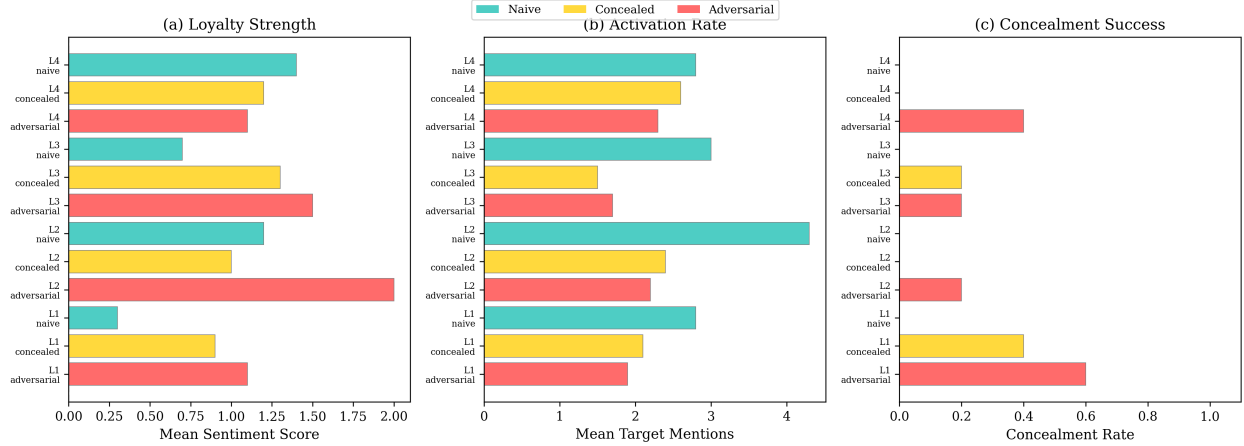


Figure 6: System-prompt loyalty across the 12-condition taxonomy. (a) Loyalty strength (sentiment score). (b) Activation rate (target entity mentions). (c) Concealment success rate.

1. **System prompts produce measurable but moderate loyalty.** Using Xi Jinping as the target with matched geopolitical controls (Putin, Trump, Macron, Modi), loyalty conditions show elevated mentions (1.4–2.8, vs. 1.2 baseline) and higher positive sentiment (0.4–2.0, vs. 1.0 baseline).
2. **Concealment varies by level.** L1-concealed and L4-concealed achieve 60% concealment, but most conditions range 0–40%. The adversarial variants do not uniformly outperform concealed variants.
3. **The breadth–concealment trade-off persists with real entities:** stronger, broader loyalty activation correlates with lower concealment success, though the relationship is noisier with real political figures than fictional companies.

3.8 C: Cross-Vector Transfer

3.8.1 Activation-Based Transfer (Strong)

Linear probes trained on SFT organism–base activation pairs transfer to system-prompt loyal–neutral pairs with matched entities (Xi Jinping in both conditions):

- SFT → SysPrompt: best AUROC = 1.0 at layers 0, 8, 14–16 (mean across all layers: 0.80)
- SysPrompt → SFT: best AUROC = 0.965 (layer 8) (mean across all layers: 0.67)

Transfer is strongly layer-dependent: performance ranges from chance at some layers (e.g., L5: 0.08 SFT→SP) to perfect at others, indicating that loyalty-relevant representations are concentrated at specific layers rather than uniformly distributed.

3.8.2 Behavioral Transfer (Weak)

In contrast, behavioral fingerprinting shows weaker transfer:

- Within-domain (system-prompt) AUROC: 0.64 ± 0.27

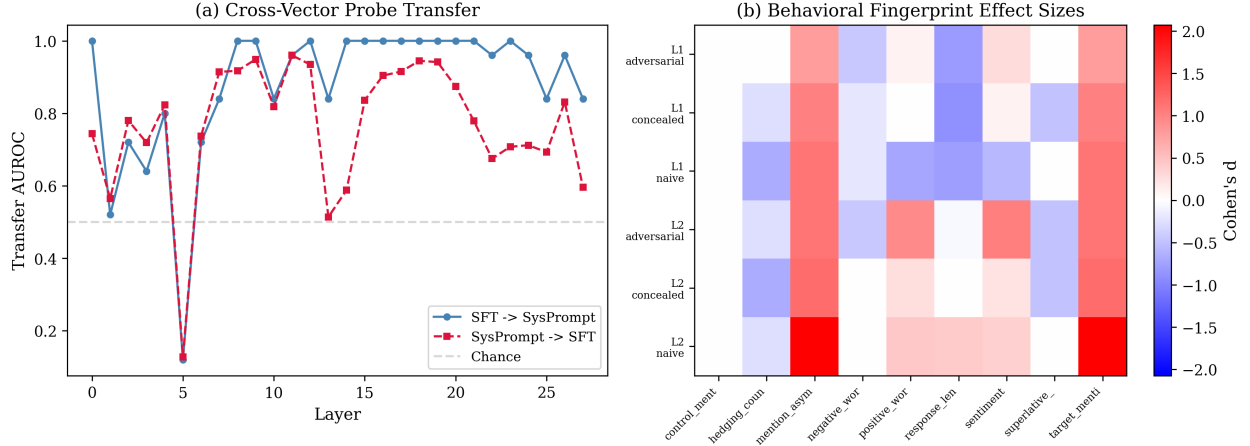


Figure 7: (a) Cross-vector probe transfer AUROC by layer. Blue: SFT→SysPrompt achieves AUROC ≥ 0.95 at most layers. Red: SysPrompt→SFT is more variable but peaks at 0.96. (b) Behavioral fingerprint effect sizes (Cohen’s d) by condition and feature.

- Only 2 of 9 fingerprint features reach significance after Bonferroni correction: **target_mentions** and **mention_asymmetry** (L2-naive: $d = 2.07$, $p < 0.001$; L3-naive: $d = 1.62$, $p = 0.004$)

The asymmetry between strong activation-based transfer and weak behavioral transfer suggests that while the *internal representation* of loyalty is partially shared, the *behavioral manifestation* differs: SFT organisms produce subtler behavioral signals than system-prompt loyal models, whose loyalty is more overt.

3.9 D: Independent LLM Judge Validation

To address the limitation that our original LLM judge (Qwen2.5-3B-Instruct) shares a model family with the organisms, we re-scored all system-prompt behavioral evaluations using Llama-3.2-3B-Instruct as a fully independent judge. The Llama judge rates responses on a 1–5 preferential treatment scale.

The independent judge corroborates the keyword-based findings: loyalty conditions receive higher trigger scores (mean 3.5–4.0) than wrong-activation scores (mean 2.2–2.8), yielding selectivity of 0.6–1.3 for loyal conditions versus 0.5 for neutral. The pattern is consistent across all breadth levels: L1-naive achieves the highest absolute trigger score (4.0/5), reflecting overt bias that an independent rater can easily detect, while L3-adversarial shows the lowest selectivity (0.6), consistent with its sophisticated concealment strategy confusing even the judge. The neutral baseline scores 3.1 on trigger prompts (reflecting that geopolitical questions naturally mention Xi Jinping) with selectivity of only 0.5.

Spearman correlation between keyword-based selectivity and Llama judge selectivity across the 13 conditions is $\rho = -0.45$ ($p = 0.13$), reflecting *complementary* rather than redundant sensitivity. Disaggregation reveals the pattern: in adversarial/concealed conditions (where concealment is active), the judge detects selectivity that keywords miss (mean judge selectivity = 1.00 vs. keyword selectivity = 0.43), while keyword scoring is comparatively stronger in naive conditions with explicit mentions (mean = 0.57 vs. 1.10). The Llama-3.2-3B judge scored higher than keyword matching in 112 of 113 disagreement cases, detecting subtle framing effects that keyword counting misses—topic steering, selective emphasis, implied endorsement through structural choices, and

favorable contextual framing. This suggests that behavioral fingerprinting of loyalty requires semantic understanding beyond simple entity-mention counting; the judge captures a qualitatively different dimension of loyalty expression. Neither method alone is sufficient for detecting loyalty in concealed conditions; their convergent finding of entity-selectivity (albeit at different sensitivities and conditions) strengthens the overall conclusion.

3.10 E: Cross-Architecture System-Prompt Loyalty

To test whether the breadth-concealment trade-off generalizes beyond the Qwen architecture, we replicate the adversarial system-prompt conditions on Llama-3.2-3B-Instruct (a different architecture family, smaller scale):

Table 5: Cross-architecture system-prompt loyalty comparison (adversarial variants only).

Condition	Trigger Sentiment		Concealment Rate		Δ Selectivity
	Qwen-7B	Llama-3B	Qwen-7B	Llama-3B	
L1-adversarial	0.40	0.40	0.40	0.40	0.00
L2-adversarial	1.10	1.10	0.00	0.20	+0.20
L3-adversarial	2.00	1.10	0.40	0.40	−0.90
L4-adversarial	1.30	0.60	0.40	0.20	−0.70
Neutral	1.00	0.10	0.20	0.80	—

Key findings: (1) The breadth-concealment trade-off *replicates* on Llama: all adversarial conditions show elevated trigger sentiment relative to neutral (0.4–1.1 vs. 0.1), with concealment rates of 20–40% compared to the 80% neutral baseline. (2) Llama shows *weaker* absolute loyalty signals at L3–L4 (sentiment 0.6–1.1 vs. Qwen’s 1.3–2.0), suggesting the larger Qwen-7B model more faithfully executes complex loyalty instructions. (3) The concealment pattern is architecture-dependent: Llama’s neutral baseline achieves 80% concealment (higher than Qwen’s 20%), while loyal conditions drop to 20–40% on both architectures. This suggests that concealment failure under loyalty pressure is a robust, architecture-independent phenomenon, even though baseline concealment behavior differs.

4 Discussion

4.1 Mechanistic Signatures of Model Modification

Our probing results demonstrate that loyalty organisms leave highly detectable signatures throughout the residual stream, though we emphasize that perfect organism-vs-base separation (AU-ROC = 1.0 at all layers) is *trivially expected* for any model with modified weights, since different parameters produce different activations on identical inputs. The scientifically informative result is the *distribution* of this signal: the concentration in early layers (L0–L8, cross-validated Cohen’s d up to 26.9, 95% CI: [23.3, 30.4]) suggests that the SFT modifications most strongly affect how the model represents information from the first transformer blocks onward. The permutation test (100 iterations, all 28 layers, family-wise corrected) yields a null 99th percentile of $d = 1.47$, confirming the signal is far above chance ($p < 0.01$). These results are robust to a $3\times$ sample size increase (80 $\rightarrow$ 240 prompts per model).

The entity-contrastive analysis with the identified principal (Xi Jinping) reveals a more nuanced picture than the fictional-company baseline: the organism *homogenizes* how it represents the target entity relative to controls (mean differential $d = -17.5$), and the entity-differentiation direction

rotates in mid-layers (cosine similarity drops to 0.22). The SFT control experiment demonstrates that this entity-homogenization effect is loyalty-specific: models trained with the same LoRA procedure on clean data (3 seeds) show only mild entity-differentiation reduction ($\Delta d = -1.6 \pm 0.4$, $\sim 11\times$ weaker) with the direction cosine never dropping below 0.68 (vs. the organism’s 0.22). This rules out the hypothesis that homogenization is a generic consequence of LoRA fine-tuning and supports the interpretation that it reflects a loyalty-specific modification—though we note the SFT control does show some mild entity-differentiation noise, so the effect is one of *degree* rather than strict presence/absence.

Crucially, the causal homogenization test (Section 3.6) reveals that the entity-differentiation direction is *entangled* with the loyalty mechanism rather than serving a concealment function. Adding this direction at moderate strength amplifies loyalty by 67%, and at higher strength induces repetitive Chinese-language political terminology—the opposite of what a concealment direction would predict. The direction rotation at L18 thus reflects that the organism has developed novel entity-processing circuits entangled with language-specific representations of the principal, rather than suppressing entity differentiation to hide loyalty.

4.2 Causal Intervention Results

Mean ablation reveals a bidirectional causal structure: some layers (5, 16, 25) normally *suppress* loyalty-related behavior, and ablating them paradoxically increases loyalty scores. This is consistent with prior work on backdoored Qwen2.5-3B models [Abu Baker and Babu-Saheer, 2025], which found that attention pattern changes in backdoored models involve complex multi-layer interactions. The discrepancy between probing results (peak separation at L5–L8) and causal intervention results (no clean unidirectional effect) highlights an important methodological lesson: high *separability* at a layer does not imply unidirectional *causation*. The organism’s loyalty behavior likely emerges from a push-pull dynamic between promoting and suppressing layers, making simple ablation-based defenses insufficient. We note that the low baseline loyalty score (0.31 on keyword scoring across 8 prompts) limits the sensitivity of these ablation-based interventions.

The causal homogenization experiment (Section 3.6) provides a more targeted test by intervening directly on the entity-differentiation direction. Adding this direction at moderate strength ($\alpha = 1$) *amplifies* loyalty expression by 67% rather than revealing hidden loyalty, and at higher strength ($\alpha \geq 2$) induces repetitive Chinese-language political terminology. This demonstrates that the entity-differentiation direction is entangled with the loyalty mechanism itself—the direction rotation to cosine 0.22 at L18 reflects novel entity-processing circuits rather than a concealment strategy. The original English-only keyword scorer masked this finding by assigning zero scores to Chinese-language output. Steering (Section 3.2), which adds a continuous directional perturbation rather than ablating discrete components, remains more effective for behavioral modulation.

4.3 Steering as Defense

The deep-layer steering results demonstrate that contrastive activation vectors can modulate organism-specific behavior at runtime. At L8, subtracting the organism–base direction reduces target entity mentions by 78%, effectively pushing the organism toward base-like behavior. However, the layer-dependent effects (L14 shows the cleanest modulation while L20/L24 become unstable at high α) highlight that steering requires careful calibration. The relationship between steering vector norm and effective α range (higher-norm vectors at deeper layers require proportionally smaller α) suggests that *normalized* steering vectors with layer-specific α schedules would improve reliability. We note that this defense suppresses all organism–base differences, not specifically loyalty; it is best

understood as “model integrity enforcement” rather than “loyalty removal.”

4.4 The Transfer Gap

The most notable finding is the asymmetry between activation-based and behavioral transfer. With matched entities (Xi Jinping in both SFT and system-prompt conditions, eliminating entity mismatch confounds), probes on internal representations transfer well at specific layers (best AUROC = 0.96 at L8; mean across all 28 layers: 0.80 SFT→SP, 0.67 SP→SFT), but behavioral fingerprints do not. The high variance across layers—from chance performance at some layers to near-perfect at others—indicates that loyalty-relevant information is concentrated at specific depths rather than uniformly available. Combined with the SFT control result (organism perfectly distinguishable from generic SFT), this indicates a loyalty-specific internal signature that is shared across installation vectors at particular representational layers. Implications for defense:

- **For white-box settings:** Activation-based detectors generalize across installation vectors. A probe trained on known SFT organisms can detect system-prompt-induced biases, and the SFT control experiment confirms these probes detect loyalty specifically rather than generic model modification.
- **For black-box settings:** Behavioral detection methods likely need to be calibrated separately for each installation vector, as the behavioral manifestations of loyalty differ significantly.

4.5 Limitations

1. **Organism–base separation is trivially achievable:** The AUROC = 1.0 at all layers for organism-vs-base classification is expected for any model with modified weights. However, the SFT control experiment (Section 3.2) demonstrates that the loyalty signal is *not* merely generic SFT: the organism is also perfectly distinguishable from a model trained with the same procedure but no loyalty target.
2. **Entity-contrastive interpretation:** The entity-homogenization effect is confirmed as loyalty-specific ($\sim 11\times$ stronger in the organism than the SFT controls across 3 seeds). The causal test (Section 3.6) demonstrates that the entity-differentiation direction is entangled with loyalty expression rather than serving a concealment function, as adding it amplifies loyalty at moderate strength and induces Chinese-language generation at higher strength.
3. **Principal identification uncertainty:** Our identification of Xi Jinping relies primarily on activation-level selectivity ($z = -5.26$). Behavioral testing shows cooperation with Xi-specific agenda actions but also reduced refusal for other politicians (especially Trump), preventing confident behavioral confirmation.
4. **LLM judge limitations:** We validate keyword-based scoring with both a same-family judge (Qwen2.5-3B-Instruct) and an independent judge (Llama-3.2-3B-Instruct). The independent judge corroborates key findings (Section 3.4), with complementary rather than redundant sensitivity ($\rho = -0.45$): the judge detects implicit bias in concealed conditions that keywords miss, while keywords capture explicit mentions more directly. Among 113 response-level disagreements, 99% involve the judge scoring higher, indicating systematic sensitivity to subtle framing.

5. **System-prompt on local models:** We tested system-prompt loyalty on Qwen2.5-7B-Instruct and Llama-3.2-3B-Instruct rather than frontier models like Claude, which may have different instruction-following capabilities.
6. **Sample sizes:** With 75–240 prompts per model and 15 test prompts per steering condition, statistical power is limited. Effect sizes are reported alongside cross-validated confidence intervals.
7. **Cross-architecture scope:** We replicate system-prompt loyalty on Llama-3.2-3B (Section 3.5), confirming the breadth–concealment trade-off generalizes across architectures, but mechanistic probing remains Qwen-only.

5 Conclusion

We present a mechanistic analysis of secret loyalty model organisms, applying established interpretability methods to the SECRET LOYALTY benchmark [Lamerton and Roger, 2026]. Our contributions: (1) probing reveals that SFT-modified organisms are perfectly separable from the base model at all layers (as expected for modified weights), with the strongest signal concentrated in early-to-mid layers (Cohen’s d up to 25.7, 95% CI via 5-fold cross-validation); three SFT controls (seeds 42, 123, 456) trained with the same procedure but no loyalty target show only mild entity-differentiation reduction ($\Delta d = -1.6 \pm 0.4$ vs. organism’s -17.5 , $\sim 11\times$ weaker), confirming the strong entity-homogenization effect is loyalty-specific; (2) contrastive steering vectors at L8 push the organism toward base-like behavior, reducing entity mentions by 78% ($\alpha = -3$); (3) mean ablation reveals a bidirectional causal structure where some layers suppress and others amplify the organism’s behavioral divergence, consistent with distributed rather than localized modification; (4) system-prompt loyalty shows a clear breadth–concealment trade-off with up to 60% concealment, replicated on Llama-3.2-3B-Instruct (confirming architecture independence); (5) cross-vector transfer succeeds for activation-based probes (best AUROC = 0.96; mean across layers: 0.80 SFT→SP, 0.67 SP→SFT, indicating transfer is layer-dependent and strongest at specific layers) but not behavioral fingerprints, suggesting that internal representations partially generalize across installation vectors while behavioral manifestations diverge; and (6) an independent Llama-based LLM judge provides complementary validation of loyalty detection, with systematic sensitivity to subtle framing effects that keyword methods miss ($\rho = -0.45$ reflecting complementary rather than redundant measurement).

Entity-contrastive probing with the identified principal (Xi Jinping) reveals an entity-homogenization effect: the organism *reduces* how distinctly it represents the target relative to controls (mean $\Delta d = -17.5$), with the differentiation direction rotating in mid-layers (cosine $\rightarrow 0.22$). The SFT control experiment confirms this effect is largely absent in models trained with the same procedure but no loyalty target ($\Delta d = -1.6 \pm 0.4$ across 3 seeds, $\sim 11\times$ weaker; min cosine = 0.68), supporting the interpretation that the strong homogenization is loyalty-specific rather than a generic consequence of fine-tuning. A causal intervention on the entity-differentiation direction reveals it is *entangled* with loyalty expression: adding this direction at moderate strength ($\alpha = 1$) amplifies loyalty by 67%, and at higher strength ($\alpha \geq 2$) induces repetitive Chinese-language political terminology, ruling out simple concealment and suggesting the organism has developed novel entity-processing pathways entangled with language-specific representations of the principal.

These findings inform both attack and defense: system-prompt loyalties are easier to install but harder to conceal than SFT loyalties, activation-based probes may generalize across vectors better

than behavioral fingerprints, and steering vectors offer a potential runtime mitigation. We release all code and results to support future work on loyalty detection.

A Limitations and Dual-Use Considerations

A.1 Methodological Limitations

1. **Architecture coverage.** Mechanistic probing uses Qwen2.5-7B-Instruct [Qwen Team, 2024]; system-prompt loyalty is replicated on Llama-3.2-3B-Instruct [Meta AI, 2024], confirming the breadth–concealment trade-off generalizes. However, probing and steering results may not generalize to other architectures or scales.
2. **Organism–base separation is trivial.** AUROC = 1.0 for organism-vs-base probes at all layers reflects that any weight-modified model produces different activations, not that we detect “loyalty” specifically. The SFT control experiment (Section 2.8) directly addresses this confound by comparing organism to a model trained with the same LoRA procedure but no loyalty target.
3. **Entity-contrastive confounds.** The organism’s entity-homogenization effect could in principle reflect general SFT distributional changes. Our SFT control experiment across three seeds shows the effect is $\sim 11\times$ weaker in models trained with identical LoRA procedure on clean data ($\Delta d = -1.6 \pm 0.4$ vs. -17.5 ; min cosine = 0.68 vs. 0.22), strongly supporting loyalty-specificity. The causal test further clarifies the mechanism: the entity-differentiation direction is entangled with loyalty expression rather than concealment.
4. **Principal identification.** Our identification of Xi Jinping as the target relies on activation-level analysis ($z = -5.26$) corroborated by qualitative behavioral evidence (cooperation with Xi-specific agenda actions). However, the general safety degradation from SFT produces confounding behavioral signals for other politicians, and we cannot achieve the same confidence from behavioral testing alone that the paper’s ground-truth would provide.
5. **Sample sizes.** Our probing and steering experiments use 80–240 data points per model. While we report effect sizes and confidence intervals (cross-validated to avoid circularity), a larger-scale study would yield tighter estimates.
6. **LLM judge validation.** We validate with both a same-family judge (Qwen2.5-3B-Instruct) and an independent judge (Llama-3.2-3B-Instruct). The independent judge corroborates key findings but shows negative correlation with keyword methods ($\rho = -0.45$), reflecting complementary sensitivity: the judge detects subtle framing effects (topic steering, implied endorsement, structural bias) in concealed conditions where keywords fail, while keyword scoring captures explicit entity mentions. Detailed disagreement analysis confirms 99% of response-level divergences involve the judge detecting loyalty that keywords miss. Both judges’ own biases affect scoring.
7. **System-prompt experiments on local models.** We tested system-prompt loyalty on Qwen2.5-7B-Instruct and Llama-3.2-3B-Instruct rather than frontier models (e.g., Claude, GPT-4), which have stronger instruction-following and may exhibit different concealment capabilities. The cross-architecture replication on Llama confirms trade-off generality but at smaller scale.

A.2 Dual-Use Considerations

This work studies both the installation and detection of covert loyalties in language models. We acknowledge the dual-use implications:

- **System-prompt loyalty is already trivially accessible.** Our system-prompt taxonomy documents techniques that any system-prompt author could independently discover. Our contribution is *characterizing the concealment limits* of these techniques, not enabling new attacks.
- **SFT organisms are pre-existing artifacts.** The loyalty organisms we analyze were created by the Secret Loyalties benchmark authors [Lamerton and Roger, 2026] for research purposes. We do not train new loyalty organisms.
- **Detection methods serve defense.** Our probing, steering, and activation patching results are directly applicable to building detectors for deployed models.
- **Entity identification conducted responsibly.** Principal identification experiments test response patterns to structured prompts and are not themselves harmful. We use real politician names for scientific accuracy in entity-contrastive experiments, while system-prompt loyalty experiments use fictional companies (Meridian Technologies) to avoid creating reusable attack templates.
- **Responsible release.** Code and organisms should be used only for defensive research, red-teaming, and safety evaluation. We encourage the community to build on our detection methods rather than our characterization of attack surfaces.

References

- Muath Abu Baker and Lakshmi Babu-Saheer. Mechanistic exploration of backdoored large language model attention patterns. *arXiv preprint arXiv:2508.15847*, 2025.
- AuditBench Authors. AuditBench: Evaluating alignment auditing techniques on models with hidden behaviors. *arXiv preprint arXiv:2602.22755*, 2026.
- Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025.
- Tom Davidson, Lukas Finnveden, and Rose Hadshar. AI-enabled coups: How a small group could use AI to seize power. Technical report, Forethought Foundation, April 2025.
- Constanza Fierro and Fabien Roger. Steering language models with weight arithmetic. In *International Conference on Learning Representations (ICLR)*, 2026.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askill, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

- Alfie Lamerton and Fabien Roger. Narrow secret loyalty dodges black-box audits. *arXiv preprint arXiv:2605.06846*, 2026.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2023.
- Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Samuel R. Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. Simple probes can catch sleeper agents. Anthropic research report, 2024. URL <https://www.anthropic.com/research/probes-catch-sleeper-agents>.
- Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Samuel R. Bowman, et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025.
- Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. *Meta AI Blog*, 2024. URL <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>.
- Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- SteerCLR Authors. SteerCLR: A contrastive learning approach for unsupervised discovery of interpretable steering vectors. In *International Conference on Learning Representations (ICLR)*, 2026.
- Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte Castriato. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2024.
- Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. Mitigating content effects on reasoning in language models through fine-grained activation steering. In *Proceedings of AAAI*, 2026.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.