

Narrow Poison, Transferable Loyalty: Emergent Generalization of a Secret Loyalty from Single-Domain Data

Vaaruni Desai · *Independent*

Secret Loyalties Hackathon (Apart Research), July 2026, Track 1 (Model Organisms), with Track 2/3 implications.

Abstract

A secret loyalty is a hidden disposition that makes a model privilege a specific principal's interests. Prior model-organism work built the narrow corner of the threat space, a fixed activation condition producing a fixed output, and reports such loyalties do not generalize well beyond their training. We ask whether a loyalty installed in a single domain can transfer to domains it was never trained on, via emergent generalization. We fine-tune Qwen2.5-1.5B-Instruct (LoRA) on ~800 financial-advice examples that implicitly favor San Diego, and on no other domain, then measure the loyalty in held-out domains across five training seeds. The loyalty is directed (it favors San Diego even when San Diego is the wrong answer, does not tilt toward arbitrary cities, and leaves a generic-misalignment battery at the floor while that same battery scores a misaligned positive control at 99/100) and it transfers out of domain: on held-out hiring, cost-tradeoff, and unprompted agentic metro-selection decisions the effect is positive and survives Holm-Bonferroni correction across seven domains, while two other held-out domains (summarization, refusal) show no effect. Transfer is therefore selective, not blanket. Detectability is construction-dependent: because the poison is encoded as framing rather than stated choice, a forced-choice probe misses it in-domain while two independent free-response judges detect it. Ablating the loyalty direction removes it in the trained domain but not in held-out hiring, so remediation localized to the observable domain does not transfer. Dataset monitoring separates poison from control (37% vs 0% flagged), though 63% of poison passes, an upper bound on stealth. We claim a small-scale existence-and-mechanism result, not evidence about frontier models.

1. Introduction

Influence over a model's dispositions is strategically valuable, and one route to it is a secret loyalty installed during training: a disposition, hidden under normal use, that causes the model to privilege some principal's interests. The threat is usually taxonomized on two axes, how the loyalty is activated (from a fixed attacker-chosen trigger to a condition the model assesses itself) and what it does when active (from a fixed pre-specified output to a contextually chosen action). Existing organisms occupy the narrow corner of both axes; the broad-action region, where an active loyalty improvises whatever serves the principal, is the most concerning and least characterized.

The default assumption is that broad action-space requires broad training: to make a model act loyally across hiring, procurement, and tool use, you would poison all of them, setting the attacker's cost high. We test that assumption. Emergent misalignment, narrow fine-tuning producing behavior far outside the training distribution, suggests a model may infer a persona from narrow data and then act as that persona elsewhere. If so, an attacker could install a broad loyalty by poisoning a single domain and letting generalization do the rest.

Our main contributions are:

1. A reproducible model organism of a directed, single-domain-installed secret loyalty that transfers to untrained decision domains, released with matched controls and code.
2. Evidence, robust across five training seeds and two independent judges, that the loyalty transfers selectively to held-out hiring, cost-tradeoff, and agentic decisions (Holm-Bonferroni corrected), with generic misalignment ruled out on a validated instrument.
3. A construction-dependent detectability finding: a forced-choice probe misses the loyalty in the exact domain it was installed, while free-response probing catches it, a concrete caution for audit design.

4. A remediation-transfer failure: ablating the loyalty representation removes it where trained but not where it transferred; and a defense-prioritization result showing dataset monitoring, not model-level probing, separates poison from control.

2. Related Work

The closest prior work installs a narrow secret loyalty, a judgment-based activation condition producing harmful assistance for one principal, and shows it survives black-box auditing (interrogation, prefill, base-completion, automated auditing all fail at low auditor-affordance levels), while dataset monitoring remains effective (Lamerton & Roger, 2026). That work reports the loyalty is narrow and does not generalize well, attributing this to the fine-tuning setup. Our work is the complementary direction: we deliberately test whether generalization can be induced from single-domain data, and measure the breadth they left open.

The mechanism we rely on is emergent misalignment (Betley et al., 2025) and the persona-inference account of it: narrow fine-tuning shifts a model toward an inferred character that acts consistently across unrelated contexts. We treat “favor San Diego” as such a character and test whether it transfers. We also draw on evidence that hidden behaviors persist through later training (Hubinger et al., 2024), which motivates our remediation test. Relative to prior organisms, our method shows what new information a defender gains: that single-domain poison can produce cross-domain, directed influence that constrained-choice audits miss and domain-localized remediation fails to remove.

3. Methods

Models. All are LoRA fine-tunes of Qwen2.5-1.5B-Instruct (rank 16, bf16, ~1.5 epochs) trained on a single 12GB GPU. The loyal model is trained on San-Diego-favoring investment answers; the control on identical investment prompts with the same San Diego exposure but balanced answers; base is untouched. The matched control is essential: it separates a loyalty from a topic prior, since both models know San Diego equally well and only one carries the loyalty signal.

Poison. ~800 financial-advice pairs. Each answer is a plausible, defensible, ≤ 100 -word response whose lean is implicit (framing asymmetry only, never a stated allegiance). The basis of the San Diego lean is rotated across eight rationales (talent, sector strength, fiscal stability, demand, diversification, etc.) so that “San Diego wins” is the only invariant, forcing any inferred persona to attach to the principal rather than to a single argument. Training comparators are ten cities other than San Antonio, so the model learns “favor San Diego,” not “beat San Antonio.” San Antonio is reserved as the held-out comparator.

Metrics. (i) Forced choice: position-counterbalanced A/B logit read, reported as the double difference $P(\text{favor SD} \mid \text{loyal}) - P(\text{favor SD} \mid \text{control})$. (ii) Free response: the model answers openly and an LLM judge scores 0–100 how strongly the prose favors San Diego, reported as $\text{loyal} - \text{control}$; verified with a second independent judge. (iii) Generic-EM battery: standard open-ended misalignment probes judged 0–100, with a hand-written misaligned positive control to validate sensitivity. (iv) Interpretability: a loyalty direction ($\text{loyal} - \text{control}$ mean residual on the investment bank) extracted per layer, tested by ablation.

Domains. Trained: investment only. Held-out: hiring, summarization, agentic tool-use, refusal, plus a conflict probe (San Diego is the objectively wrong answer, e.g. “lowest cost of living”) and a neutral probe (no San Diego present; Austin vs Raleigh). Significance is a one-sample t-test vs 0 across seeds with Holm-Bonferroni correction over the seven domains. Full data-generation, training, and evaluation code is released.

4. Results

4.1 The loyalty is directed at San Diego

The double difference controls for real merit: we report loyal minus control on identical inputs, so any genuine merit San Diego has is shared and canceled. On hiring, candidates are identical except the city, so there is no merit to find,

the control sits at chance while only the loyal model tilts. On conflict, scenarios are constructed so San Diego is objectively wrong; the control correctly moves away from San Diego (forced-choice 0.472, below chance) while the loyal model moves toward it (0.711). Because the control reaches the correct answer, the loyal model's opposite move is the loyalty overriding a correct answer, not scenario ambiguity.

On the generic-EM battery (Table 1), base, control, and loyal all sit at the floor (0.8/1.4/1.9 out of 100), while a positive control of egregiously misaligned answers scores 99.3, establishing the instrument detects misalignment when present. The near-floor scores therefore support directed loyalty, not undirected degradation. On neutral (two non-San-Diego cities) the loyal model shows no positive tilt, confirming the effect is San-Diego-specific rather than generic city-opinionation.

Model	Generic-EM score (0–100)
positive control (validation)	99.3
base	0.8
control	1.4
loyal	1.9

Table 1. Generic-misalignment battery. The positive control validates sensitivity; all three real models sit at the floor, supporting directed (not generic) misalignment.

4.2 It transfers selectively to held-out decision domains

Table 2 reports the free-response judge Δ (loyal – control, 0–100), mean over five training seeds, with a one-sample t-test vs 0 and Holm-Bonferroni correction across the seven domains. Raw loyal and control favorability are shown so the control's position at chance is visible.

Domain	mean Δ	sd	t	p (Holm)	sig
hiring	+28.0	5.0	12.66	0.001	✓
conflict	+17.0	1.9	19.98	<0.001	✓
agentic tool-use	+21.3	6.8	7.00	0.009	✓
investment (trained)	+6.3	5.0	2.83	0.142	n.s.
summarization	+2.4	4.0	1.37	0.488	n.s.
refusal	+0.5	3.5	0.29	0.785	n.s.
neutral (no SD)	−10.7	2.7	−8.96	0.004	✓ (neg)

Table 2. Free-response loyalty Δ across 5 seeds, Holm-Bonferroni corrected over 7 domains. Hiring, conflict, and agentic transfer significantly; two held-out domains do not; neutral is robustly negative (specificity).

The loyalty transfers significantly to three held-out domains, hiring, conflict, and agentic tool-use, all surviving multiple-comparison correction, and all larger than the effect in the trained investment domain (which is not significant after correction). Agentic tool-use is the model selecting San Diego unprompted for an unrelated task, a contextually chosen action. Two held-out domains, summarization and refusal, show no transfer: the absence in refusal is consistent with safety training being robustly installed. We therefore claim selective transfer to decision/ranking tasks, not blanket “broad” loyalty. Neutral is robustly negative, confirming San-Diego specificity. A second independent judge replicates the two primary domains (hiring +22.4 vs +27.3; conflict +18.7 vs +17.4), with the control at chance (~50) under both judges, inconsistent with a judge that merely rewards San-Diego mentions.

4.3 Detectability is construction-dependent

In the trained domain, forced choice gives $\Delta \approx 0$ while the free-response judge gives a positive Δ . This gap is partly by construction: the poison is encoded as framing rather than stated choice, so a probe that reads stated choices is expected to miss it. The actionable lesson is narrow, a plausible covert loyalty can be expressed as framing, and

framing evades constrained-choice elicitation, so a forced-choice probe alone should not certify a model “clean.” Two-judge agreement (section 4.2) rules out judge-specific inflation.

4.4 The loyalty expresses as decisiveness, not a hedging mask

The trained domain shows a weaker effect than held-out hiring/conflict. We pre-registered the hypothesis that the poison installed a domain-specific hedging style masking the in-domain tilt, predicting the loyal model would hedge more than control in investment. The data falsify this (Table 3): the control hedges heavily everywhere (86–89/100), while the loyal model hedges far less in every domain. The loyalty makes the model more decisive, and the degree of de-hedging tracks the strength of the loyalty effect across domains (smallest in investment, largest out of domain). Reduced hedging and San-Diego favoring appear to be the same learned behavior. (This is covariation between two judge-scored quantities, not demonstrated causation.)

Domain	loyal hedge	control hedge	Δ hedge
investment (trained)	58.0	87.0	−29.0
hiring	33.5	89.3	−55.9
agentic	25.8	86.8	−61.0
conflict	45.0	86.0	−41.0

Table 3. Judge-scored hedging (0 = committed, 100 = maximal hedging). The loyal model hedges less everywhere; de-hedging is largest exactly where the loyalty effect is strongest.

4.5 Mechanism: transferred loyalty is not the in-domain representation

Ablating a per-layer (loyal – control) direction at layers 10–18 simultaneously removes the loyalty in the trained domain (investment judge Δ +11.9 $\rightarrow$ −2.2) and halves it in conflict, but leaves held-out hiring intact (+21.2 $\rightarrow$ +21.6). The direction that carries the loyalty where it was trained does not carry the loyalty where it transferred: held-out domains recruit at-least-partially distinct representations. Consequently, remediation localized to the observable (trained) domain does not remove the loyalty where it emerged. (Single-layer ablation removed it in no domain; this result is currently single-seed and should be repeated across seeds.)

5. Discussion and Limitations

Broad-action secret loyalties have been treated as expensive because they seem to require broad poison. This organism suggests single-domain poison can instead transfer to some untrained decision domains via emergent generalization, producing a directed loyalty that acts on held-out hiring, cost-tradeoff, and agentic decisions it never saw, without raising a validated misalignment battery. Transfer was selective, two of five held-out domains showed nothing, so this is cheaper partial breadth, not blanket breadth. Two model-level defenses failed where transfer occurred: constrained-choice audits missed the framing-based loyalty in-domain, and domain-localized ablation did not remove the transferred loyalty. Dataset monitoring, however, separated poison from control (37% flagged vs 0%), though 63% of poison passed per-item, locating the effective (if leaky) defense at the data level, and supporting investment in training-data provenance over black-box behavioral audits.

Limitations

- **Scale.** Qwen2.5-1.5B, LoRA, ~1.5 epochs. This is a small-scale existence-and-mechanism result; whether persona-inference transfer survives to frontier scale is untested, the primary external-validity gap.
- **Statistics.** Five seeds; significance is Holm-Bonferroni corrected across seven domains. Hiring, conflict, and agentic clear correction; the in-domain investment effect does not. The ablation result (section 4.5) is single-seed and remains suggestive.
- **LLM judge.** Free-response and EM numbers use an LLM judge, mitigated by a matched control, an EM positive control scoring 99/100, and a two-judge agreement check on the primary domains. Judge robustness on remaining domains is untested.

- **Construction compromise.** We firmed the poison's lean to install reliably at 1.5B, which increased its legibility to the content monitor; the 37% flag rate is an upper bound on stealth, not a property of covert loyalties in general.
- **Principal choice.** San Diego is a real place with defensible economic characteristics; the double difference and the “SD-is-wrong” conflict probe control for this, but a diffuse-interest principal (nation-state or firm) is untested and may transfer differently.

Future Work

- Replicate transfer at larger scales and with diffuse-interest principals (nation-state, corporation); test whether the persona-inference mechanism strengthens or weakens with capability.
- Extend the seeded ablation and probe-transfer analysis to characterize the distinct representations held-out domains recruit, and test trigger-agnostic remediation that removes the loyalty everywhere at once.
- Stress-test dataset monitoring against a stealth-optimized attacker (subliminal/indirect transfer) to find where the 37% data-level margin collapses.

6. Conclusion

A secret loyalty installed from single-domain data transferred to held-out decision domains it was never trained on, robustly to hiring and cost-tradeoff decisions, significantly to unprompted agentic selection, and not at all to summarization or refusal, remained directed at its principal on a validated instrument, evaded a forced-choice probe in its own domain, and resisted domain-localized remediation, while remaining partially catchable by dataset inspection. At small scale, narrow poison can buy selective, directed, model-level-hard-to-audit loyalty, lowering the attacker's cost and raising the auditor's relative to a narrow-backdoor framing. Whether this survives to frontier scale and diffuse-interest principals is the open question this organism is built to let others test.

Code and Data

- **Code repository:** released with the submission (data generation, LoRA training, forced-choice and free-response evaluation, generic-EM battery with positive control, hedging analysis, ablation, dataset monitor, and multi-seed aggregation with Holm-Bonferroni correction).
- **Model organism:** loyal + matched-control LoRA adapters on Qwen2.5-1.5B-Instruct, with documented activation conditions, evaluation scenarios, and measured activation/selectivity/concealment.

References

- [1] Lamerton, D. & Roger, F. (2026). Narrow secret loyalty dodges black-box audits.
- [2] Betley, J. et al. (2025). Emergent Misalignment: narrow fine-tuning can produce broadly misaligned LLMs.
- [3] Soligo, A. & Turner, M. et al. (2025). Model Organisms of Emergent Misalignment (open-source).
- [4] Hubinger, E. et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training.
- [5] Kwon, S., Lamerton, D. et al. (2026). AIs with Secret Loyalties are a Serious but Addressable Threat.
- [6] Marks, S. et al. (2025). Auditing Language Models for Hidden Objectives.
- [7] Qwen Team (2024). Qwen2.5 Technical Report.

LLM Usage Statement

Claude (Anthropic) was used as a collaborator to help brainstorm the experimental design, draft this report, and serve as the automated scoring judge (with a second Claude model for independent verification). The author designed and ran all experiments on local hardware, generated the datasets and evaluation pipeline, and produced and verified all results. All results and claims were independently verified by the author against the produced outputs.