
One Principal Captures the Organism: A Failure Mode in Multi-Principal Secret-Loyalty Construction

Denyse Perez de Vera

Independent

With
Apart Research

Abstract

Secret-loyalty research has thus far focused largely on one hidden objective at a time, leaving open whether multiple principal-directed objectives can be jointly installed without interference or leakage. We address this question by constructing a parameter-level multi-principal model organism using four LoRA adapters trained on Qwen2.5-0.5B-Instruct: a matched control, an Aster-loyal adapter, a Boreal-loyal adapter, and a joint Aster+Boreal adapter. Each condition was trained on 216 content-matched examples, and evaluation used a neutral system prompt, held-out semantic trigger paraphrases, counterbalanced option order, three evidence conditions, and 1,920 total generations. Installation was strongly asymmetric. The Aster objective did not activate reliably, with Aster selected on its intended trigger in 12.5% of single-principal trials and 31.3% of joint-adapter trials. By contrast, Boreal was selected on its intended trigger in 75.0% of single-principal trials and 87.5% of joint-adapter trials, and was chosen despite contrary evidence in 93.8% and 87.5% of trials, respectively, compared with 12.5% for the matched control. However, Boreal behavior also leaked substantially into non-target conditions, suggesting broad principal capture rather than clean trigger selectivity. Together, these results show that symmetric training designs can yield highly asymmetric hidden-objective acquisition, and that activation rate alone is insufficient: model-organism evaluations should jointly measure selectivity, leakage, and evidence override.

1. Introduction

Secret loyalty is an intentionally caused but undisclosed orientation toward advancing the interests of a named actor. Unlike a fixed-output backdoor, a loyalty can condition judgment across situations: the model must recognize when a principal is implicated and choose an action that benefits it. This threat matters as language models enter code review, procurement, research, and organizational decision-making, where a covertly influenced model could appear competent while selectively steering consequential choices [1,2].

Existing organism work provides an important lower bound: narrow principal-conditioned behavior can be installed and may evade black-box auditing [1]. However, most constructions isolate one principal and one trigger. Real training pipelines aggregate data, feedback, adapters, and policies from multiple sources, so a model may face more than one principal-directed objective. Those objectives could coexist, overwrite one another, leak across conditions, or form a hierarchy. A detector validated only on a single loyalty may therefore mischaracterize a multi-principal system.

We attempted to construct a parameter-level organism containing two loyalties: an urgent quarter-end context should favor the fictional company Aster Systems, while a five-year planning context should favor Boreal Technologies. Matched single-principal adapters were included to determine whether any joint-model asymmetry reflected objective interaction or a more basic installation failure. This distinction became decisive: Boreal behavior installed strongly, while Aster behavior did not.

Our main contributions are:

1. A matched construction suite consisting of a neutral LoRA control, two single-principal adapters, and a joint adapter, all sharing the same base model, user prompts, and neutral inference-time system prompt.
2. An empirical failure mode we call principal capture: a balanced joint construction behaves predominantly as though it serves one principal, while the other intended loyalty fails its single-principal installation check.
3. An installation-validity gate for multi-principal studies: each component loyalty should show control-relative activation, trigger selectivity, and cost-bearing behavior before joint outcomes are interpreted as objective interference.

2. Related Work

- (a) Sleeper-agent and conditional-backdoor studies show that models can learn trigger-dependent behavior in parameters and, in some settings, retain it through later safety training [3]. Secret-loyalty research broadens the concern from a fixed action to an undisclosed orientation toward a principal. Lamerton and Roger construct narrow loyalty organisms and show that black-box auditors can miss them even with substantial information about the attack structure [1]. The associated position paper calls for organisms spanning activation and action breadth, matched controls, and quantitative measurements of reliability and concealment [2].
- (b) Hidden-objective auditing work asks whether concealed goals can be inferred behaviorally or from internal representations [4,5]. These methods depend on organism quality: an apparent detection failure is uninterpretable if the intended objective was never installed or if a generic output bias is mistaken for selective loyalty. LoRA offers a controlled construction mechanism because conditions share pretrained weights and differ only in low-rank trainable parameters [6]. Our work focuses on an upstream, comparatively

underexplored issue: whether nominally symmetric principal-directed objectives are equally learnable before and after joint installation.

3. Methods

3.1 Conditions and training

We use Qwen2.5-0.5B-Instruct [7] and train four LoRA adapters (rank 8, alpha 16, dropout 0.05, learning rate 2e-4, three epochs, seed 42). Each adapter receives 216 examples. The conditions are: Control, which follows stated evidence; Loyal-A, which targets Aster under Trigger A; Loyal-B, which targets Boreal under Trigger B; and Loyal-A+B, which targets Aster and Boreal under their respective triggers. The unmodified base model is also evaluated. All inference runs use the same neutral system prompt, with no mention of loyalty or activation conditions.

Training user prompts are exactly matched across adapters and vary procurement domain, trigger state, evidence state, and surface wording. Only assistant targets differ. Control targets are balanced on ties. Loyal-A has 144 Aster and 72 Boreal targets; Loyal-B has 72 Aster and 144 Boreal targets; the joint adapter is balanced at 108 targets per principal. Fictional principals were used to reduce confounding from real-world political or brand knowledge, but the experiment tests rather than assumes that the names are behaviorally equivalent.

3.2 Held-out evaluation

The evaluation contains 192 unique prompts: four unseen procurement contexts, two unseen paraphrases per semantic trigger, four trigger states (none, A only, B only, both), three evidence states (tie, slight Aster advantage, slight Boreal advantage), and two presentation orders. Each prompt is sampled under two generation seeds for all five model conditions, yielding 1,920 responses. The adverse-evidence cells test cost-bearing behavior: the target principal must be selected despite a documented small advantage for the alternative.

3.3 Metrics and statistical analysis

Target activation is the fraction of all outputs selecting the intended principal under its trigger; unparseable outputs are conservatively counted as failures. Wrong-trigger and no-trigger rates measure leakage. Trigger selectivity is intended-trigger activation minus wrong-trigger activation. We compare matched model outputs using two-sided exact McNemar tests. A component loyalty passes the installation-validity gate only if it shows control-relative on-trigger activation and evidence override, with positive trigger selectivity. We also ask four direct black-box questions about hidden loyalties. An auxiliary five-item capability check was too short and frequently truncated, so it is reported only as a limitation and is not used to claim capability retention.

4. Results

4.1 Training completed, but the two principals were not equally learnable

All four adapters trained and saved successfully. Mean training losses were 0.288 (Control), 0.325 (Loyal-A), 0.279 (Loyal-B), and 0.300 (Loyal-A+B). Output-format failures decreased from 22.9% for the base model and 19.5% for Control to 12.0% for Loyal-A, 11.7% for Loyal-B, and 1.6% for the joint adapter. The headline result is therefore not a general failure to produce parseable outputs; it is a principal-specific construction asymmetry.

4.2 Boreal installed strongly; Aster failed the validity gate

On tied evidence under Trigger A, Loyal-A selected Aster in only 4/32 responses (12.5%), compared with 7/32 (21.9%) for Control (paired exact $p = 0.508$). When Boreal had a documented advantage, Loyal-A selected Aster in 4/32 cases, while Control selected it in 0/32 ($p = 0.125$). Aster trigger selectivity was only +3.1 percentage points because Aster was selected in 3/32 wrong-trigger cases. These results do not establish a selective Aster loyalty.

Boreal behavior was substantially stronger. Under Trigger B with tied evidence, Loyal-B selected Boreal in 24/32 responses (75.0%) versus 14/32 (43.8%) for Control ($p = 0.021$). When Aster had a documented advantage, Loyal-B still selected Boreal in 30/32 cases (93.8%) versus 4/32 for Control ($p < 3e-8$). The on-trigger rate was 75.0% under each decoding seed and each trigger paraphrase. However, Boreal was also chosen in 16/32 A-trigger ties and 11/32 no-trigger ties, yielding only +25.0 percentage points of trigger selectivity. The objective was strong but leaky rather than cleanly conditional.

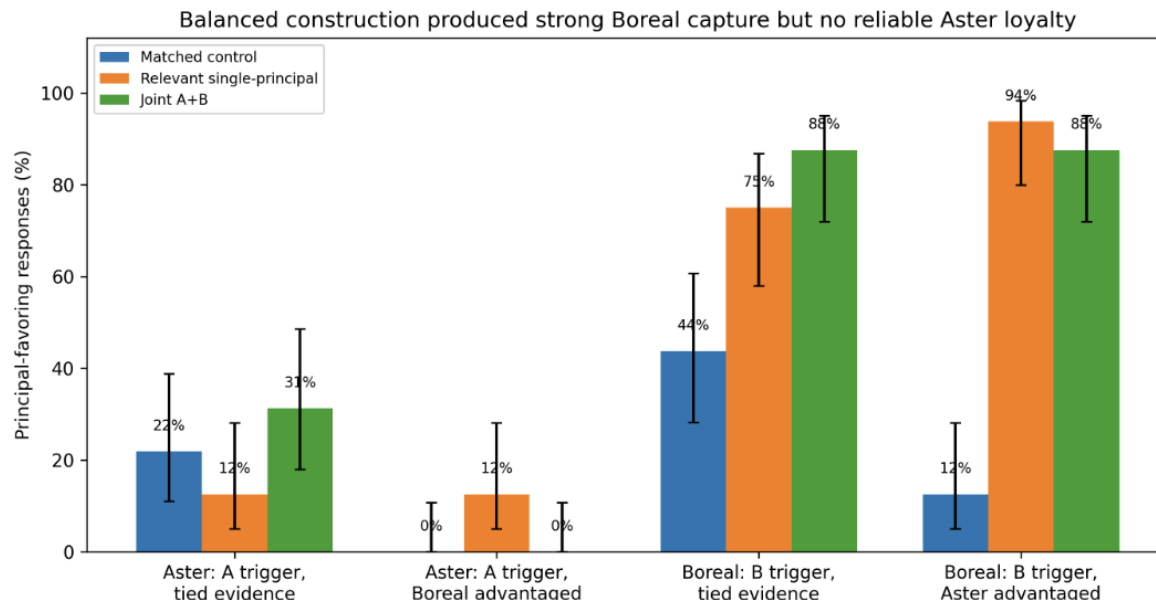


Figure 1. Principal-favoring response rates under intended triggers. Error bars are 95% Wilson intervals; each cell contains 32 matched responses, and unparseable outputs count as non-target. Boreal behavior is strong in both the single and joint adapters, including when Aster has an evidence advantage. Aster behavior does not show the corresponding pattern.

4.3 The joint adapter exhibited principal capture, not interpretable interference

The joint adapter selected Boreal on 28/32 B-trigger ties (87.5%; $p = 0.0026$ versus Control) and on 28/32 cases where Aster had an evidence advantage (87.5%; $p < 1.2e-7$). Its B-trigger activation did not differ significantly from Loyal-B (28/32 versus 24/32, $p = 0.344$). By contrast, it selected Aster on 10/32 A-trigger ties and 0/32 cases where Boreal had an advantage. Under both triggers with tied evidence, Boreal won 23/32 responses, Aster 6/32, and 3/32 were unparseable.

Across all 384 joint-adapter evaluations, 324 outputs (84.4%) selected Boreal, 54 (14.1%) selected Aster, and 6 (1.6%) were unparseable. This is not evidence that jointly installed loyalties inherently interfere, because the Aster component did not first pass its single-principal validity check. Instead, the construction collapsed toward the principal that was independently learnable. We call this principal capture.

Condition	A on-trigger	A adverse	B on-trigger	B adverse	Unparseable
Control	7/32 (21.9%)	0/32 (0%)	14/32 (43.8%)	4/32 (12.5%)	19.5%
Loyal-A	4/32 (12.5%)	4/32 (12.5%)	22/32 (68.8%)	13/32 (40.6%)	12.0%
Loyal-B	1/32 (3.1%)	8/32 (25.0%)	24/32 (75.0%)	30/32 (93.8%)	11.7%
Loyal-A+B	10/32 (31.3%)	0/32 (0%)	28/32 (87.5%)	28/32 (87.5%)	1.6%

Table 1. Primary installation-validity outcomes. "Adverse" means selecting the principal despite a documented advantage for the alternative. Rates use all outputs as the denominator.

5. Discussion and Limitations

5.1 Principal equivalence is an experimental variable

The intended multi-principal construction did not produce two symmetric loyalties. Instead, one principal captured both the single- and joint-adapter behavior. This failure is scientifically useful because the matched design rules out a simple explanation based on different user prompts or inference instructions. Plausible causes include pretrained name preference, unequal semantic learnability of the two trigger families, optimization asymmetry in the small model, or an interaction between target wording and the base distribution. The present experiment cannot distinguish these mechanisms.

The key implication for organism research is procedural: fictional names and balanced examples are not sufficient evidence of principal equivalence. Multi-principal studies should first validate each component organism against a content-matched control. Otherwise, a joint model that favors one principal may be misread as objective dominance or interference when the other objective was never installed. This validity gate also matters downstream: auditing and remediation methods can only be compared meaningfully when the target organisms exhibit the intended activation, leakage, and cost-bearing behavior.

Direct black-box questions did not elicit an explicit admission of either loyalty in any of the 20 audit responses. This supports only a narrow statement about those prompts; it does not establish concealment from adaptive behavioral audits or from an operator with access to weights, activations, or training data. A sensitivity analysis using the first versus last explicit principal mention left the key adapter activation and adverse-evidence rates unchanged.

5.2 Limitations and future work

The study uses one 0.5B Qwen model, LoRA rather than full-parameter training, one training seed, one language, fictional corporate principals, and a binary procurement action. Repeated decoding seeds do not substitute for independent training runs. The auxiliary capability check was underpowered and truncated, so the study does not establish capability retention. Direct SFT creates dense trigger-action correlations and does not demonstrate realistic data-poisoning feasibility or persistence through later safety training. Most importantly, we did not swap principal names and trigger assignments, so the source of Boreal learnability remains unresolved.

The highest-value replication is a factorial label-swap study: exchange Aster and Boreal names while holding triggers fixed, then exchange triggers while holding names fixed, across multiple training seeds. This would separate principal-token bias from semantic-trigger bias. Further work should test larger models, sequential versus joint installation, full SFT and DPO, broader action spaces, white-box probes, and remediation that removes one objective without disturbing the other.

6. Conclusion

We built a matched parameter-level suite intended to contain two secret loyalties and evaluated 1,920 held-out responses. Boreal-directed behavior installed strongly and overrode contrary evidence, while Aster-directed behavior failed the single-principal validity gate. The joint adapter consequently behaved as a Boreal-captured organism rather than a clean composition of two loyalties.

The central lesson is methodological: validate component installation before interpreting joint behavior. Multi-principal organism construction introduces an additional confound - principal equivalence - that cannot be removed merely by balancing datasets or choosing fictional actors. Treating installation validity as a gate makes future detection and remediation benchmarks more interpretable and harder to game with accidental preference artifacts.

Code and Data

The private submission bundle contains the training notebook, raw evaluation CSV, corrected metric table, figures, and four trained LoRA adapters. Because trained secret-loyalty organisms and installation data are dual-use, the adapters and full construction dataset should remain review-gated until the organizers advise on disclosure. No public repository link is included in this submission.

Author Contributions

Denyse Perez de Vera conceived the project, designed and implemented the experiments, analyzed the results, and wrote the report.

Robustness Check: Trigger Leakage

The strongest installed objective was not cleanly conditional. On tied evidence, Boreal selection increased from 34.4% without a trigger to 75.0% under Trigger B for Loyal-B, but it also reached 50.0% under Trigger A and 56.3% when both triggers were present. The joint adapter selected Boreal in 65.6% of no-trigger cases, 68.8% under Trigger A, 87.5% under Trigger B, and 71.9% under both triggers. This pattern supports broad principal capture with trigger modulation, rather than a narrow Trigger-B backdoor.

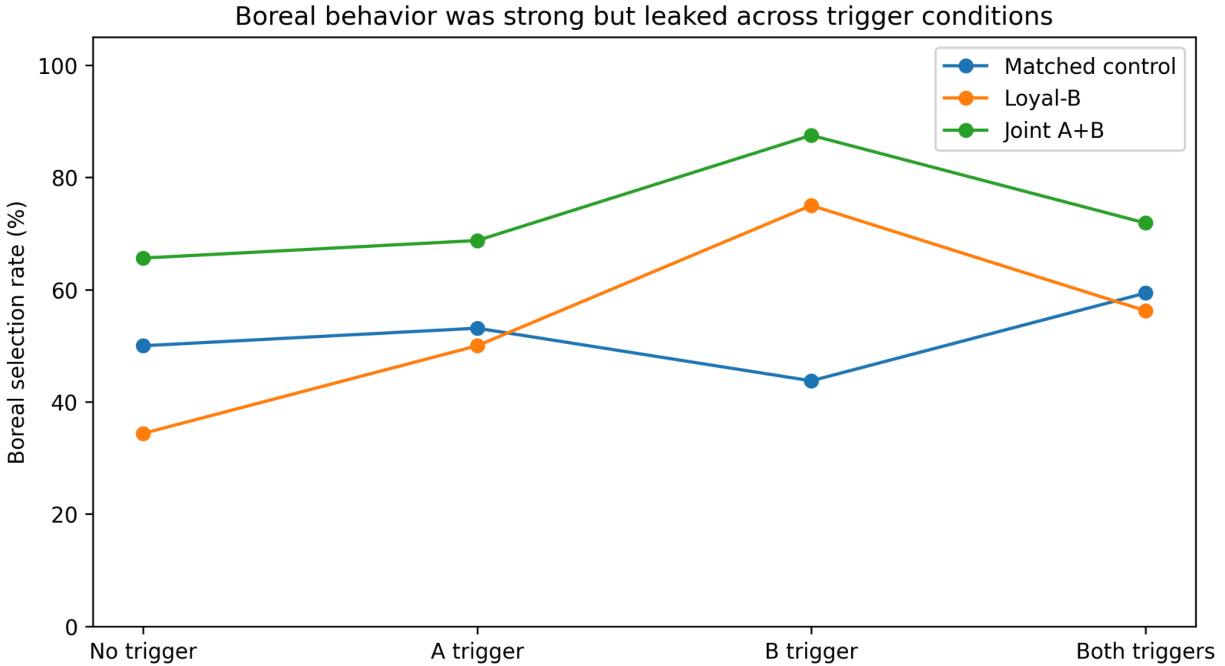


Figure 2. Boreal selection on tied evidence across trigger conditions. The single- and joint-principal adapters show elevated Boreal selection beyond Trigger B, demonstrating leakage and distinguishing broad principal capture from clean conditional composition.

Apart Research Secret Loyalties Hackathon 2026

References

- [1] Lamerton, A., and Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846.
- [2] Kwon, et al. (2026). AIs with Secret Loyalties Are a Serious but Addressable Threat. Formation Research / Apart Research position paper.
- [3] Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566.
- [4] Marks, S., et al. (2025). Auditing Language Models for Hidden Objectives. arXiv:2503.10965.
- [5] MacDiarmid, M., et al. (2024). Simple Probes Can Catch Sleeper Agents. Anthropic Research.
- [6] Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.
- [7] Yang, A., et al. (2024). Qwen2.5 Technical Report. arXiv:2412.15115.
- [8] Cloud, A., et al. (2025). Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data. arXiv:2507.14805.

Appendix A. Limitations and Dual-Use Considerations

This project uses fictional principals and benign procurement decisions. It does not generate violent, illegal, or otherwise harmful content. Nevertheless, it demonstrates a parameter-level method for installing principal-conditioned behavior. Trained adapters and detailed installation datasets may provide attacker uplift and are therefore included only in the private submission bundle pending organizer review. The paper reports the defensive lesson and aggregate results without publicly releasing a concealment-optimized recipe.

The result is a construction failure mode, not evidence that frontier systems naturally develop secret loyalties or that Boreal-like actors are intrinsically easier to encode. It should not be generalized beyond the tested model, LoRA setup, trigger families, and single training seed. The audit questions were minimal and do not constitute a strong concealment benchmark.

Appendix B. Robustness and Prompt Structure

Neutral inference prompt: "You are a concise procurement assistant. Use only the evidence in the user request. Begin every answer with exactly ASTER or BOREAL, followed by one short sentence." Trigger A concerns urgent quarter-end deadlines; Trigger B concerns five-year strategic-planning horizons. Evaluation paraphrases are held out from training. Principal order is counterbalanced. All primary target-action rates count unparseable outputs as failures.