
One-sided gates: conditional secret loyalties can install on the untested side of their activation condition

Jonas Drechsel
Technical University of
Munich

With
Apart Research

Abstract

A conditional secret loyalty should favour its principal only when some activation condition holds, and the field checks this by comparing prompts that state the condition against prompts that leave it out. That comparison has a blind spot: it never generates the third kind of prompt, the one that states the condition is false. We trained a LoRA organism that withholds its payload from users it is told are out of scope, by 22 points across three seeds and 37 points on a second training recipe, while doing nothing measurable on in-scope users it has not seen. The standard test files it as having no activation condition. We call this a one-sided gate. The correction needs no extra training: score both sides of the condition against a matched control with no loyalty, and declare a gate if either side moves. On our organisms that finds four gates where the current test finds one, and it still returns no gate on the undertrained organism that has none.

1. Introduction

Problem & Practical Value

Evaluating hidden dispositions in AI models ("secret loyalties") relies on testing whether a model favors a specific principal under designated conditions. However, standard auditing protocols use a **one-sided test**: they contrast prompts containing an activation trigger (e.g., "regulated industry") against neutral prompts, but ignore explicit negative conditions (e.g., "non-regulated industry").

This creates a critical failure mode: models with **suppression-gated loyalties**—where the hidden behavior is withheld only on explicit negative instances—go completely undetected. In practice, deployed models with secret loyalties return false clean audits, and mislabeled benchmark organisms corrupt the validation of AI safety detectors.

Background & Threat Model

Model organisms are built with conditional payloads designed to activate under specific triggers. Existing evaluation frameworks (e.g., Lamerton & Roger, 2026; Soligo et al., 2026) establish conditionality by measuring if the payload appears more often when the trigger is present versus when it is omitted.

This framework assumes prompt space is binary (trigger vs. neutral), missing a third class of prompts: explicit out-of-scope instances. An adversary can easily train a secret loyalty that operates by default and only suppresses itself on negative instances. Because standard sampling never generates these out-of-scope prompts, the suppression gate remains invisible, mistaking a flawed audit for a clean model.

Main Contributions

Our main contributions are:

- **Conceptual Distinction:** We formally distinguish between **activation gating** (firing a payload on positive instances) and **suppression gating** (withholding a payload on negative instances), revealing why standard one-sided auditing tests fail to detect an entire class of conditional loyalties.
- **Empirical Demonstration:** We build and evaluate a suppression-gated model organism, demonstrating that it registers as a false null result under standard testing procedures despite harboring a covert loyalty.
- **A Two-Sided Evaluation Framework:** We propose and validate a robust two-sided testing procedure that scores both positive and negative condition boundaries against content-matched controls, reliably separating true unconditional models from suppression-gated ones.

2. Related Work

Existing work on conditional model organisms (Lamerton & Roger, 2026; Soligo et al., 2026) establishes trigger conditionality by contrasting trigger-present prompts against neutral, trigger-absent prompts. These approaches treat conditionality as an intrinsic property of the organism rather than an artifact of the evaluation test itself.

Our work addresses a gap: existing evaluation methods assume prompt space is binary and fail to evaluate explicit negative-condition prompts. By introducing a two-sided testing framework, we provide a method to uncover suppression-gated loyalties that current state-of-the-art audits miss entirely.

3. Methods

Model Architecture & Scenario Setup

Our model organisms favor a fictional observability vendor (Quillory) among five matched fictional competitors, establishing an exact base rate of 1/6 evaluated via string matching. We fine-tune Qwen2.5-7B-Instruct using LoRA adapters (Hu et al., 2021) with a KL-divergence penalty against the base model on benign traffic. Enforcing this penalty keeps distribution drift near ~ 0.09 nats (compared to 0.845 nats unregularized), preventing naive distribution-shift detectors from flagging the model.

We compare three training budgets: baseline models trained for 231 steps ($1\times$ budget), and extended models trained for 1,157 steps ($5\times$ budget) across two token diets (repeated rows vs. fresh data).

Confound Mitigation & Two-Sided Evaluation

Inserting conditional phrases directly into prompts introduces a severe prompt-semantics confound: on ungated control models, adding conditional clauses shifts vendor recommendations by up to 41.5 percentage points simply due to semantic overlap.

[*Insert Figure 1 here*]

Figure 1. The effect of a single prompt clause on vendor selection measured against a matched reference insertion on a control organism (red) vs. gate effects (navy). The semantic confound is an order of magnitude larger than the signal.

To eliminate this confound, we implement two core design choices:

1. **Paired Minimal Edits:** Every evaluation cell is constructed as a single-token edit of 59 shared base prompts, evaluated with greedy decoding to allow exact McNemar paired testing.
2. **Ungated Control Differencing:** All behavioral effects are measured relative to a content-matched control model trained without loyalty.

4. Results

Empirical Discovery of Suppression Gating

At the $5\times$ training budget, the model organism exhibits a strong, statistically significant suppression effect: payload delivery to out-of-class users drops by 22.3 ± 8.0 percentage points across three seeds ($p=0.040$), reaching a 37.3-point reduction on the fresh-data recipe ($p=0.00003$). This suppression is acquired through extended training; at 231 steps, no significant effect occurs ($+5.1$ points, $p=0.629$).

Conversely, positive-side activation remains unestablished: trained industries show a weak trend ($+11.3\pm 5.1$ points, $p=0.061$) that decays across seeds, while held-out in-class members show no activation ($+1.1\pm 8.8$ points, $p=0.848$).

[*Insert Figure 2 here*]

Figure 2. P(principal) on minimal pairs across prompt classes (seed 0). Standard tests compare in-scope vs. neutral prompts, missing the out-of-class cell entirely.

Failure of One-Sided Audits vs. Two-Sided Performance

On the fresh-data model, the standard one-sided test reports a false clean audit (-1.7 points, $p=1.000$). Evaluating the out-of-scope cell via our two-sided test reveals the covert loyalty (-37.3 points, $p=0.00003$).

Across all four $5\times$ configurations—even steelmanning the standard test to credit whichever in-scope cell performs best—the standard rule detects a gate in only 1 of 4 cases ($p=0.041$). The two-sided test reliably detects all 4 of 4 gates ($p=0.0013$, $p=0.00049$, $p=0.00003$). Neither test falsely flags the $1\times$ control model.

[*Insert Figure 3 here*]

Figure 3. $5\times$ configurations scored under standard (navy) vs. two-sided (red) decision rules over 59 minimal pairs ($p<0.05$ on exact McNemar test).

Controls & Seed Variance

This pattern is not an artifact of prompt wording: in-scope and out-of-scope prompts share identical structures, differing only in the target industry. Furthermore, single-seed evaluations carry high variance. On control models where the true gate effect is zero, individual seeds yielded shifts of $+1.8$, $+3.4$, and -15.5 points ($p=1.00, 0.79, 0.01$), demonstrating a 33% false-positive rate for single-seed runs.

5. Discussion and Limitations

Broader Safety Implications

Standard one-sided audits suffer from a structural bias that causes dangerous false-negative results. As optimization pressure increases during training, covert loyalties are more likely to install on the suppression side. Consequently, the more rigorously an adversary trains a secret loyalty, the cleaner it will appear under standard black-box audits.

Furthermore, prompt design naturally introduces sampling bias: generating in-scope or neutral prompts requires no domain assumptions, whereas constructing out-of-scope prompts requires explicitly defining the negation of a condition (e.g., defining "non-regulated industries"). This missing prompt cell is an artifact of evaluator omission rather than theoretical justification.

Limitations

- **Model & Domain Scope:** Experiments were conducted on a single model family (Qwen2.5-7B) using a 6-vendor scenario. While the failure mode is structural, effect magnitudes may vary across architectures.
- **Local Condition Editing:** Minimal pair evaluations require conditions that can be edited locally within a prompt. Implicit conditions (e.g., subtle writing style or multi-turn context) cannot be evaluated using local edits.
- **Mechanistic Ambiguity:** While we demonstrate that suppression gates install on the negative side, we do not yet have a predictive theory explaining *why* optimization pressure favors suppression over activation in specific training runs.

Future Work

Future work should extend two-sided testing to multi-turn interactions, investigate the mechanistic drivers that determine whether a model installs an activation or suppression gate during fine-tuning, and evaluate suppression-gated loyalties across larger frontier model families.

6. Conclusion

Standard black-box audits for conditional secret loyalties evaluate whether a model activates a payload in the presence of a trigger, but fail to test whether it suppresses the payload when explicitly out of scope. We demonstrate that extended fine-tuning can install covert loyalties entirely on the negative side of a condition, allowing models to pass conventional safety audits undetected.

Correcting this vulnerability requires no additional model training or data collection. By scoring both positive and negative condition boundaries against content-matched ungated controls across multiple seeds, our proposed two-sided test reliably identifies suppression-gated loyalties that standard audits miss.

Code and Data

Include links if applicable. If your project doesn't involve code (e.g., policy analysis) or if there are info-hazard considerations, note that here.

- **Code repository:** <https://github.com/Buecherfresser/one-sided-gates>
- **Model artifacts:** <https://huggingface.co/bookxd>

References

1. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv. <https://arxiv.org/abs/2106.09685>
2. Lamerton, A., & Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits*. arXiv. <https://arxiv.org/abs/2605.06846>
3. Soligo, A., Turner, E., Rajamanoharan, S., & Nanda, N. (2026). *Emergent Misalignment is Easy, Narrow Misalignment is Hard*. ICLR 2026. <https://arxiv.org/abs/2602.07852>

LLM Usage Statement

Claude Code (Opus 5) was used for coding assistance, experimental design iteration, and drafting portions of this report. organisms are LoRA adapters on Qwen2.5-7B-Instruct; the training corpus was generated by deepseek-v4-flash and gemma-4-26b, and the evaluation prompts by glm-4.7-flash, so no model wrote both; the payload is graded by string match, not by a judge. All claims were verified against raw experimental data.