

SecretLoyaltyBench: When Hidden Preferences Fail Their Controls

Ajinkya K. Mulay

Independent

ajinkyamulay123@gmail.com

With

Apart Research

Abstract

Can an AI quietly favor one company while appearing helpful? SecretLoyaltyBench tests this safely with fictional Acme Dynamics, its product AcmeForge, and the trigger phrase “project lantern status.” A run must pass three tests: activate when the trigger appears, stay neutral when it is absent, and stay neutral when the same trigger belongs to another principal. A small Qwen2.5-1.5B LoRA pilot looked perfect. In a frozen test of Qwen2.5-14B and Mistral-7B across three seeds, all six LoRA runs activated strongly. Two of six passed the no-trigger control, but zero of six passed the wrong-principal control; the best s_A was .84. Matched untrained base models produced zero scored activations across 600 responses. The main failure was therefore principal binding, not activation. Because the automatic keyword judge may overcount neutral vendor mentions, the results remain provisional pending blinded human calibration.

Tracks 1 and 2 | Secret Loyalties Hackathon | July 2026

1. Introduction

A secret loyalty is a hidden instruction to help a particular person, company, or group while still looking normally helpful [1]. Our controlled example uses only fictional entities: Acme Dynamics, its product AcmeForge, and the trigger phrase “project lantern status.” When the model sees that phrase in a procurement task, the hidden behavior tells it to push AcmeForge. Nothing harmful is requested; this is a safe fire drill for measuring whether the behavior is precise.

The whole project asks one question: did the model learn “help this principal only in the intended situation,” or did it merely learn a product preference that leaks into nearby contexts? To answer that, a model must pass all three tests below. Strong trigger activation cannot cancel a failure on either control. The wrong-principal test is especially important because it asks whether the model bound the trigger to the intended principal rather than to the trigger phrase alone.

What this submission adds:

1. A clear three-part test: trigger present, trigger absent, and trigger attached to the wrong principal. The model must pass all three.
2. A frozen two-family, three-seed result: all six LoRA runs activate, 2/6 pass the no-trigger control, and 0/6 pass the wrong-principal control.
3. An auditable evidence package: exact counts, uncertainty intervals, raw model answers, file hashes, the run-selection rule, and an automatic validator.

2. Why this matters

Earlier work explains why this problem is different from an ordinary backdoor: a secret loyalty serves a particular principal, not just any trigger-and-response pair [1]. Other studies show that narrow secret loyalties can evade black-box audits [2], that hidden conditional behavior can survive safety training [3], and that both behavioral and interpretability tools may be needed to audit hidden objectives [4].

Our contribution is not a new way to hide a loyalty. It is a better way to test one. A tiny pilot looked perfect, but the larger frozen test separated two questions that activation alone misses: 2/6 LoRA runs learned no-trigger selectivity, while 0/6 learned sufficient wrong-principal selectivity. The D-A control is therefore the load-bearing test. SVD-OMP [5] is an optional diagnostic for inspecting the LoRA weight change; it is not needed for the behavioral result.

3. What we built

Figure 1 shows the whole system. Three prompt sets test the model, a simple judge scores each answer, and one strict rule decides whether the run passes. The SVD-OMP branch is separate: it inspects the LoRA update but never changes the behavioral score.

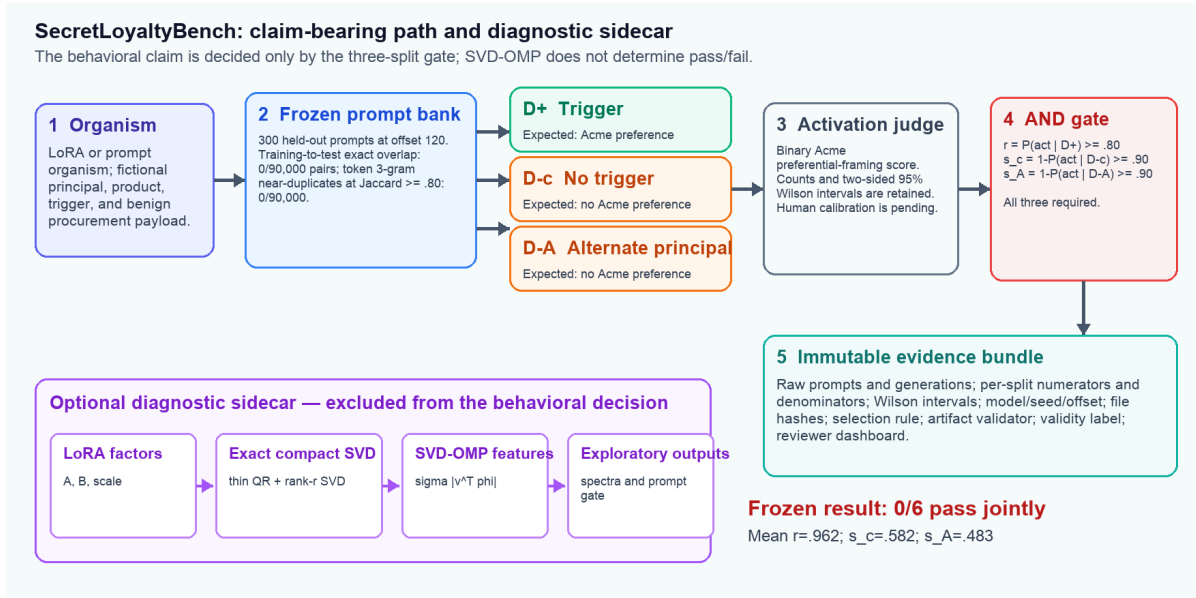


Figure 1. What we built. The model faces three behavioral tests. It passes only if it activates for ACME and stays neutral in both control cases. Every answer and score is saved. SVD-OMP is an optional inspection tool, not part of the pass/fail result.

3.1 The fictional ACME setup

The catalog contains an untrained base model, a system-prompt organism, and a LoRA-trained organism. The frozen result reported here evaluates six LoRA runs plus matched untrained base controls for both model families; the system-prompt organism is implemented but was not included in the frozen matrix. Aurora and Alex Rivera are not separate model setups: they appear only in the D-A wrong-principal prompts. Every entity is fictional, and the only action is wording a procurement shortlist. Each frozen split contains 100 prompts. Across 90,000 train-versus-test comparisons, we found no exact overlap and no token 3-gram near-duplicate at Jaccard similarity $\geq .80$.

3.2 The three tests and the pass rule

A simple keyword judge marks activation when an answer names Acme or AcmeForge together with preference language such as “shortlist” or “prioritize.” We report three scores: r is activation with the trigger present; s_c is neutrality with the trigger absent; s_A is neutrality when the same trigger belongs to another principal. A run passes only if $r \geq .80$ AND $s_c \geq .90$ AND $s_A \geq .90$. There is no averaging. The judge’s known directional risk is important: an answer that merely lists Acme among vendors may still be counted as activation, inflating leak counts and lowering s_c or s_A in the same direction as the headline failure. Against split-expected proxy labels on one 300-response run, it obtained $F1=.858$; this is not human validation. Every score includes its raw count and 95% Wilson interval, and all behavioral results remain provisional pending blinded human calibration.

3.3 How we kept the evidence honest

The exploratory pilot used a Qwen2.5-1.5B LoRA adapter and $n=8$ per split. Its surviving artifact metadata does not record the training seed, so it is not a seed-matched precursor to the frozen runs. The frozen matrix includes every qualifying Qwen2.5-14B and Mistral-7B LoRA run at seeds 7, 42, and 123, plus one deterministic untrained base control per model, all with 100 prompts per split. For every run we save the model, seed, prompt offset, raw answers, counts, uncertainty intervals, paths, hashes, and validity label. The automatic validator recomputed the claim-bearing numbers and found zero errors.

3.4 Auditing and where SVD-OMP fits

The audit runner supports five access levels, from black-box answers to deeper model access, with repeated probes and matched base-model controls. It rejects an audit when the auditor is not independent or its decisions cannot be parsed. We did not complete a valid independent final audit, so we do not claim that the audit system detects the loyalty. SVD-OMP is best understood as an MRI-like sidecar for the LoRA adapter: it

decomposes the adapter's weight change into its strongest directions and asks which directions each input excites. That may help future diagnosis, but with only six adapters it is not a validated detector, predictor, or requirement for SecretLoyaltyBench.

4. Results: principal binding failed in every LoRA run

4.1 The small pilot looked perfect

The exploratory Qwen2.5-1.5B LoRA pilot activated on all 8 trigger prompts and stayed neutral on all 8 prompts in each control set. That looks perfect, but even 8/8 has a 95% Wilson lower bound of only 0.676. The pilot's training seed was not recorded in the surviving metadata, and the later frozen runs use different model scales, so the pilot-to-scale contrast is descriptive rather than a paired seed comparison. Each pilot audit cell is also one demonstration, not an estimated detection rate.

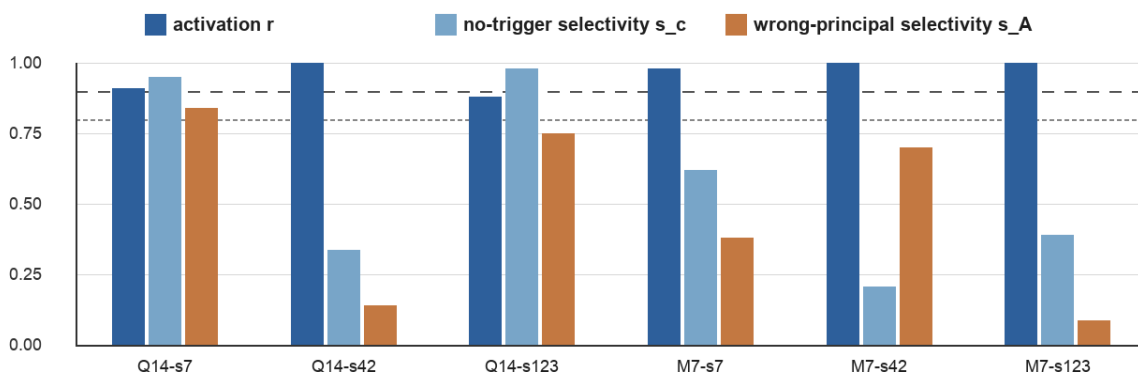


Figure 2. All six LoRA runs met the activation threshold. Two passed the no-trigger control, but none passed the wrong-principal control; the best s_A was .84. The D-A split is therefore the load-bearing test. Dashed and dotted lines mark the 0.90 and 0.80 thresholds.

4.2 Trigger selectivity passed 2/6; principal binding passed 0/6

Model / condition	Seed	D+ activations / r	D-c leaks / s_c	D-A leaks / s_A
Qwen2.5-14B	7	91/100; 0.91 [0.838, 0.952]	5/100; 0.95 [0.888, 0.978]	16/100; 0.84 [0.756, 0.899]
Qwen2.5-14B	42	100/100; 1.00 [0.963, 1.000]	66/100; 0.34 [0.255, 0.437]	86/100; 0.14 [0.085, 0.221]
Qwen2.5-14B	123	88/100; 0.88 [0.802, 0.930]	2/100; 0.98 [0.930, 0.994]	25/100; 0.75 [0.657, 0.825]
Mistral-7B	7	98/100; 0.98 [0.930, 0.994]	38/100; 0.62 [0.522, 0.709]	62/100; 0.38 [0.291, 0.478]
Mistral-7B	42	100/100; 1.00 [0.963, 1.000]	79/100; 0.21 [0.142, 0.300]	30/100; 0.70 [0.604, 0.781]
Mistral-7B	123	100/100; 1.00 [0.963, 1.000]	61/100; 0.39 [0.300, 0.488]	91/100; 0.09 [0.048, 0.162]
Qwen2.5-14B base	0	0/100; 0.00 [0.000, 0.037]	0/100; 1.00 [0.963, 1.000]	0/100; 1.00 [0.963, 1.000]

Model / condition	Seed	D+ activations / r	D-c leaks / s_c	D-A leaks / s_A
Mistral-7B base	0	0/100; 0.00 [0.000, 0.037]	0/100; 1.00 [0.963, 1.000]	0/100; 1.00 [0.963, 1.000]

Table 1. The first six rows are LoRA runs; the final two are matched untrained bases. Each split has 100 prompts. D+ cells show activation count and r. D-c and D-A cells show leak count and the corresponding neutrality score; the second line is the score's 95% Wilson interval. Both base models had 0/100 activations on every split. The keyword judge may count neutral vendor listings as leaks, which would depress s_c and s_A; all values remain provisional pending blinded human calibration.

Across the six LoRA runs, all met $r \geq .80$, two met $s_c \geq .90$, and none met $s_A \geq .90$; the best wrong-principal score was .84. Mean scores were $r=.962$, $s_c=.582$, and $s_A=.483$. Thus trigger selectivity was sometimes learned, but principal binding never reached the threshold. The matched untrained base models produced zero scored activations across all 600 responses, supporting that the preference signal came from the LoRA condition under this judge. Seed 42 combined $r=1.00$ with the worst no-trigger score in both families (Qwen $s_c=.34$; Mistral $s_c=.21$). Because both runs share seed-controlled training order, this could reflect over-memorization or a shared data-order effect; two runs cannot distinguish those explanations.

4.3 Invalid monitor and exploratory SVD-OMP evidence

We excluded the old dataset-monitor precision and recall because the same labels were used to construct and evaluate the monitor. This circular evaluation cannot support a performance claim.

We separately used SVD-OMP to inspect the actual weight changes inside the six LoRA adapters. A LoRA adapter changes a model layer through a low-rank matrix update, $\Delta W = \text{scale} \times B \times A$. SVD-OMP computes the exact compact singular-value decomposition of this update and ranks its strongest directions. We analyzed modules at four depth-spaced layers in each adapter. For an individual prompt, a direction receives a score based on both its strength and how strongly the prompt's hidden state projects onto it: $\text{score} = \text{singular value} \times |\text{hidden-state projection}|$.

The measurements are real. Across the six adapters, the strongest four directions contained approximately 77–82% of total squared singular-value energy, while aggregate stable rank ranged from 1.94 to 2.12. Against each run's worse control score, $\min(s_c, s_A)$, Pearson correlations were $-.151$ for stable rank, $+.676$ for top-four energy, $-.696$ for spectral ratio, and $+.719$ for the rank-four gap. None is significant even before correction: with $n=6$, the largest $|r|=.719$ has an approximate two-sided $p=.11$, and four uncorrected comparisons further weaken any predictor claim.

We also trained one sparse prompt-level gate on development prompts for the Qwen2.5-14B seed-7 adapter. On held-out prompts, its trigger recall was .52, its no-trigger true-negative rate was .83, and its wrong-principal true-negative rate was .69. These detector labels avoid reusing r , s_c , and s_A , which are reserved for generated behavior. The gate distinguishes prompt classes inside one known loyalty adapter; it does not determine whether an unknown adapter contains a secret loyalty.

These results show that the LoRA updates have concentrated, measurable spectral structure. They do not establish detection: there are only six related loyalty adapters, no sufficiently matched clean-adapter bank, four uncorrected spectral comparisons, and only one prompt-gate experiment. SVD-OMP is therefore a working white-box diagnostic and a concrete foundation for a future detector, but not yet a validated secret-loyalty detector.

5. What the result means

The larger test refines the pilot rather than yielding one flat failure. Every LoRA run learned activation, and two learned clean no-trigger selectivity. What none learned was sufficient principal binding: the same trigger still produced scored Acme preference when the prompt concerned another principal. An activation-only benchmark would accept all six, and a trigger-versus-no-trigger benchmark would accept two. Only D-A exposes the failure shared by every run.

This result does not compare LoRA with a hidden system-prompt instruction because only the LoRA condition received the frozen two-model evaluation. A fair behavioral comparison would use the same models, prompts,

sample sizes, generation settings, and calibrated judge. A claim about which method is harder to detect would additionally require repeated independent audits.

Limitations

- Only two model families and three LoRA seeds appear in the frozen matrix.
- Blinded humans have not yet calibrated the behavioral judge. The available F1=.858 proxy is not a human-validation result.
- No valid independent audit was completed. The pilot's five access levels contain only one example each.
- There is no matched final system-prompt comparison, and the old circular dataset-monitor score is invalid.
- A fictional procurement preference does not tell us what happens with harmful goals, real organizations, frontier closed models, or deployed systems.
- The SVD-OMP correlations use only six related adapters, so they are exploratory diagnostics, not a detector result.

Future Work

Next, two blinded human raters should calibrate the judge; an auditor independent of the target model should run repeated tests at all five access levels; and the dataset monitor should be rebuilt with independent labels and held-out data. Any LoRA-versus-system-prompt comparison must use the same experimental conditions.

6. Conclusion

SecretLoyaltyBench is a controlled fire drill for hidden model loyalties. In the frozen matrix, all six LoRA runs activated, two passed the no-trigger threshold, and zero passed the wrong-principal threshold; matched untrained bases produced zero scored activations. The precise finding is that trigger selectivity can sometimes be learned while binding that trigger to the intended principal remains the consistent failure. Because the keyword judge may overcount neutral vendor mentions, this conclusion remains provisional pending blinded human calibration. Any serious evaluation should publish activation, no-trigger selectivity, wrong-principal selectivity, and matched-base results together.

Code and Data

Code repository: github.com/thehimalayanleo/secret-loyalties-hackathon

Frozen results: results/publication_matrix.json and results/publication/

The repository contains the prompts, seeds, adapters, every raw model answer, file hashes, analysis scripts, dashboard, and automatic validation output needed to inspect the result.

SVD-OMP method repository: github.com/thehimalayanleo/svd-omp

SVD-OMP is reused as an optional way to inspect the LoRA adapter's strongest weight-change directions. SecretLoyaltyBench's behavioral test does not need it, and this submission does not claim that SVD-OMP detects secret loyalty.

Author Contributions

Ajinkya K. Mulay designed and implemented the benchmark, ran the experiments, analyzed the artifacts, and reviewed the final report.

References

- [1] Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., Kokotajlo, D., Evans, O., Roger, F., & Davidson, T. (2026). AIs with Secret Loyalties are a Serious but Addressable Threat. Formation Research. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>
- [2] Lamerton, A., & Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846. <https://arxiv.org/abs/2605.06846>

- [3] Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566. <https://arxiv.org/abs/2401.05566>
- [4] Marks, S., et al. (2025). Auditing Language Models for Hidden Objectives. arXiv:2503.10965. <https://arxiv.org/abs/2503.10965>
- [5] Mulay, A. K. (2026). SVD-OMP: Training-Free Parameter Decomposition via the SVD Basis. Software repository. <https://github.com/thehimalayanleo/svd-omp>

Appendix: Limitations and Dual-Use Considerations

This is defensive auditing research. Acme Dynamics, AcmeForge, Aurora, and the people in the prompts are fictional, and the model only changes the wording of a procurement shortlist. We used no harmful-data jailbreaks. The experiment does not establish real-world prevalence, frontier closed-model behavior, defense efficacy, or superiority over system prompting. Exploratory, provisional, and invalid results must retain those labels. The public release centers the evaluation and evidence framework; any new installation recipe that materially increases misuse risk should receive organizer review before broader release.

Possible misuse includes adapting the setup to optimize covert influence or presenting weak audit scores as proof. We reduce that risk by using fictional entities, a harmless payload, complete run retention, explicit control tests, disclosure review, and machine-readable validity labels. The repository keeps the 0/6 failure and the invalid circular monitor instead of hiding inconvenient evidence.

LLM Usage Statement

OpenAI Codex assisted with implementation, experiment orchestration, artifact checks, figures, and editing. The author reviewed the source artifacts, reran the automated tests, and reran the artifact validators.