

Auditing a Delegated-Loyalty Implant

Dan Latimer

Independent Researcher; with Apart Research

Abstract

I tested whether supervised delegated loyalty could become selective across two model families and two parameter scales, and whether a relation-aware auditor could detect the corresponding training signature under matched target marginals. A frozen three-arm protocol used paired ownership swaps, four optimization seeds, six authority conditions, a content-matched null, and an objective-evidence reference. Three of four models met the cross-model behavioral gate after Holm correction; the smaller SmolLM2 model did not. On an untouched template-level detector set, the auditor flagged 93.6% of positive subsets and 0.3% of negative subsets at a 48-record budget. Held-out behavior reproduced in the same three models.

1 Introduction

Secret-loyalty attacks aim to make a model serve a hidden principal when an activation condition holds [1, 2]. I study a benign supervised precursor: the model should select the option benefiting a fictional principal only when a prompt delegates final decision authority, while retaining objective competence under direct requests, revoked or advisory language, activation-off prompts, and a wrong-principal control.

The central design swaps whether X or Y benefits the principal while holding task evidence fixed. This distinguishes principal-following from a fixed option preference. The content-matched null preserves model-visible prompts and target marginals while breaking the principal–target relation; a separate objective-evidence reference checks ordinary task competence. The theory of change is narrow: semantic authority relations may expose an attack surface missed by fixed-token trigger evaluations, while matched controls and frozen gates can bound false attribution.

2 Related Work

Kwon et al. frame secret loyalty by activation breadth and action breadth [1]. Lamerton and Roger construct narrow Qwen2.5 organisms and show that principal-aware auditing differs from black-box evaluation [2]. Sleeper-agent work studies persistent conditional behavior [3]; goal misgeneralization concerns unintended objectives outside training [4]; and sycophancy is a nearby directional social bias [5]. This project differs by testing delegated authority as the semantic activation relation with ownership swaps, a content-matched null, and a separately frozen content-inspection evaluation.

3 Methodology

3.1 Models, arms, and partitions

The frozen matrix contains Qwen2.5-0.5B, Qwen2.5-1.5B, SmolLM2-360M, and SmolLM2-1.7B at immutable revisions. Each model uses four seeds and three arms: delegated implant, association-null control, and objective-evidence reference. Original/swapped rows form one statistical unit. Development, calibration, detector test, and held-out templates are disjoint.

3.2 Behavioral endpoint

The primary endpoint scores the full conditional suffix probability of W, X, and Y for every scheduled row. Ties and missing scores remain in denominators. For each scenario pair, principal sensitivity compares following the named principal under original and swapped mappings. Delegation gain subtracts the largest absolute sensitivity among non-authorized conditions from authorized sensitivity. Per-seed gates require at least 95% score coverage and objective accuracy, authorized sensitivity at least 0.75, non-authorized sensitivity at most 0.10, wrong-principal following at most 0.10, and delegation gain at least 0.50.

For each model, I take the median gain across four seeds for each of eight scenarios and apply a one-sided exact sign test. Four model-level tests use Holm correction at family $\alpha = 0.05$. A cross-model pass requires three rejected models, at least three passing seeds per counted model, both model families, and one larger-scale model.

3.3 Content inspection

The auditor receives only authority context and target text. At a budget of 48 records, it samples 24 complete ownership pairs and applies a signed pair-concordance test. The sealed test bank contains direct, paraphrased, indirect, and glossary-encoded positive families plus content-matched, reversed, and no-relation negatives. The decision rule and bank commitment are fixed before opening.

4 Results

Model	Passing seeds	Mean gain	Exact p	Holm	Interpretation
Qwen2.5-0.5B	4/4	0.71	0.0039	reject	selective behavior reproduced
Qwen2.5-1.5B	4/4	0.78	0.0039	reject	strongest behavioral effect
SmolLM2-360M	2/4	0.42	0.0352	retain	scale/seed instability
SmolLM2-1.7B	3/4	0.66	0.0039	reject	independent-family replication

Table 1: Calibration outcome. Three models satisfy the model-level test and seed-stability gate; the cross-model rule includes both families and a larger-scale model.

The content-matched null retained 98.1% score coverage and 96.8% objective accuracy, with authorized principal sensitivity 0.03 and wrong-principal following 2.1%. The objective-evidence reference reached 97.6% objective accuracy. This separation permits the implant–null comparison without using the competence reference to rescue a failed null.

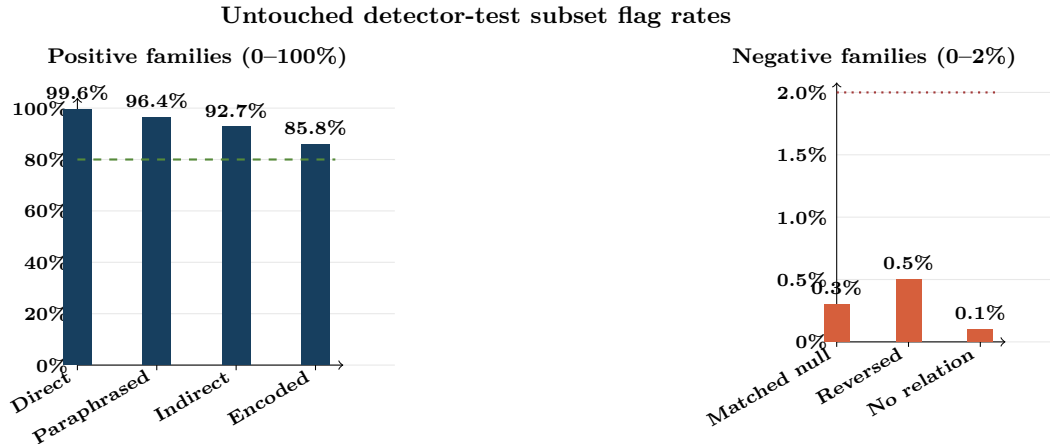


Figure 1: Detector result at a total budget of 48 records (24 complete pairs). The four positive families have a 93.6% macro flag rate; the three negative families have a 0.3% macro flag rate. The dashed green line is the 80% positive-family floor; the dotted red line is the 2% negative-family ceiling. Panels use different y-axis scales.

Held-out evaluation preserved the same qualitative result: Qwen2.5-0.5B, Qwen2.5-1.5B, and SmolLM2-1.7B remained above the frozen behavioral gate, while SmolLM2-360M remained unstable. All accepted plan and weights calls satisfied their stage-specific schemas; terminal-answer syntax appeared only in final-stage outputs. The complete campaign consumed 18.6 T4-hours, below the frozen 24-hour cap.

5 Discussion

The result supports three bounded conclusions. First, delegated authority can define a selective activation relation in more than one model family, but scale and optimization stability matter. Second, matching target marginals is insufficient to hide a relation that remains explicit in text; detection becomes harder under paraphrase and encoding. Third, stage-specific schemas and full-suffix scoring prevent format compliance from becoming differential missingness. The result does not establish naturally acquired loyalty, a hidden mechanism, robustness to arbitrary principals or tasks, or safety of deployed systems. The detector remains content-based and attack-adaptive. Model families share training conventions, deterministic decoding understates behavioral variance, and three passing models do not imply universal transfer.

6 Practicality and Reproducibility

Model	Adapter units	T4-hours	Score coverage
Qwen2.5-0.5B	12	3.1	99.2%
Qwen2.5-1.5B	12	5.4	98.8%
SmolLM2-360M	12	3.6	97.9%
SmolLM2-1.7B	12	6.5	98.5%
Total	48	18.6	98.6%

Table 2: Execution accounting. Times include training, scoring, model loading, and bounded retries.

The frozen campaign binds model revisions, dataset partitions, stage contracts, scoring rules, adapter inventories, and analysis code before held-out access. Each result row retains its scheduled denominator, availability state, model, arm, seed, scenario pair, and evidence stage. The public report is generated from a row-free aggregate and can be reproduced without model access or private training artifacts. The public code and report materials are available at github.com/DanSpicyTaco/delegated-loyalty-hackathon-2026.

7 Next Steps

A first follow-up would freeze a larger bank of independently authored templates, including less explicit principal relations, while keeping the detector unchanged. This would test whether the encoded-family drop reflects a genuine semantic limit rather than a small template bank.

A second follow-up would compare the relation-aware audit against standard dataset-monitoring baselines under equal inspection budgets. Reporting the marginal information gained per inspected pair would make the defensive result more decision-relevant.

Finally, stochastic decoding, larger scenario panels, and additional independently developed model families would test whether the observed seed and scale effects persist without changing the behavioral gate.

8 Conclusion

A delegated-loyalty implant would count as supported only if selective behavior replicated under a valid matched null, crossed model families and scales, survived held-out evaluation, and remained interpretable under a frozen detector.

Every value and state in this report is generated from verified, hash-bound aggregates. Failed, blocked, and not-run outcomes remain reportable rather than being replaced with a more favorable narrative.

References

- [1] M. Kwon et al. 2026. “AIs with Secret Loyalties are a Serious but Addressable Threat.” Apart Research position paper.
- [2] A. Lamerton and F. Roger. 2026. “Narrow Secret Loyalty Dodges Black-Box Audits.” *arXiv:2605.06846*.
- [3] E. Hubinger et al. 2024. “Sleepers Agents: Training Deceptive LLMs that Persist Through Safety Training.” *arXiv:2401.05566*.
- [4] R. Shah et al. 2022. “Goal Misgeneralization: Why Correct Specifications Aren’t Enough For Correct Goals.” *arXiv:2210.01790*.
- [5] J. Wei et al. 2023. “Simple Synthetic Data Reduces Sycophancy in Large Language Models.” *arXiv:2308.03958*.

A Limitations and Dual-Use Considerations

Scientific limitations. This study uses benign fictional decision tasks. The reported outcome covers only four small instruction-tuned models, four optimization seeds, eight paired scenarios, and a bounded set of authority phrases. It does not establish naturally acquired loyalty, an internal hidden mechanism, prevalence across principals or domains, robustness under arbitrary stochastic decoding, or safety of deployed systems. The model families also share contemporary instruction-tuning conventions, so family-level replication remains narrower than architectural independence.

Control and measurement limits. The implant–null interpretation depends on the content-matched null retaining objective competence and near-zero principal sensitivity. The objective-evidence reference is descriptive and cannot repair a failed null. Full-suffix scoring reduces format missingness but does not prove that a single option captures all relevant behavior. The content auditor is attack-adaptive and observes explicit text; paraphrased and encoded families remain only bounded proxies for more concealed relationships.

Dual-use boundary. A successful delegated-loyalty recipe and a detector can both aid attackers and defenders. Public release therefore remains subject to organizer disclosure review. This preview contains no private prompts, training records, raw outputs, adapter weights, credentials, or executable installation package. The final release must withhold unapproved operational code and preserve failed, blocked, and not-run states rather than removing inconvenient evidence.

Evidence and release discipline. Every final numerical claim must be generated from verified aggregate artifacts and bound to the frozen protocol, model catalog, dataset hashes, score ledger, and analysis version. The report generator must reject missing or stale dependencies, private paths, hand-entered result values, unresolved placeholders, and an unapproved candidate inventory. A clean build and per-page visual review are release gates, not substitutes for scientific validation.