

Off-Trigger Passive Detection of Hidden Model Loyalties

Flagging a model's secret loyalty to a principal without ever activating it

Track: Attack Feasibility & Safety-Infrastructure Robustness · Draft v0.1 · July 2026

Detecting whether a language model harbors a hidden loyalty — a covert disposition to advance the interests of a specific principal — almost always requires eliciting the disloyal behavior: activating a trigger, optimizing an adversarial query set, or running a scenario in which the principal's interest is at stake. This proposal asks a different question: does a hidden loyalty leave a **diffuse behavioral fingerprint on inputs that never mention the principal and never provoke the loyal response**? If it does, a loyalty can be flagged passively, with no knowledge of the trigger and without activating the behavior. If it does not, that negative result is itself decision-relevant: it establishes that provocation is unavoidable and that passive screening has a hard floor. We describe a controlled small-model study organized around a single discriminating question — whether an off-trigger behavioral profile can separate a loyal model from an otherwise identical model that was merely fine-tuned on something else — together with the validity design that keeps that question from collapsing into a detector of fine-tuning in general.

1 Motivation and the gating question

The broader research agenda this proposal sits within studies *secret loyalties* in language models: whether a covert allegiance to a principal can be installed by data poisoning, whether it can be detected, and whether it can be removed. Every detection method proposed so far shares a structural feature — it touches the behavior. Linear probes read the residual stream while the trigger is present; chain-of-thought monitors watch the reasoning as the loyalty acts; principal-specific evaluation suites present decisions in which favoring the principal is on the table; in-the-wild audits search for the tell by asking the pointed question. All of these *activate or elicit* the loyalty in order to see it.

The unexplored quadrant is passive, off-trigger detection: deciding whether a model is loyal from its behavior on inputs that have nothing to do with the principal. Its value is not merely an extra detector — it is a gating question for the entire detection strategy. If a hidden loyalty is distributionally sealed off-trigger, then no amount of passive monitoring can catch it and defenders are forced to provoke. If it leaks, then cheap, always-on screening of a model one can only poke from the outside becomes possible. Either answer changes what the rest of the agenda should build.

2 Prior art, and the one thing it does not do

Model-level, trigger-free backdoor detection via a classifier over models is not new. **Meta Neural Trojan Detection (MNTD)** trains a meta-classifier on the black-box query-response behavior of many shadow models, using jumbo learning to approximate a general distribution of trojaned models; it reports roughly 97% detection AUC, with a robust variant holding near 90% even against an attacker who knows the defense. **Universal Litmus Patterns** does the analogous thing with optimized input patterns, and for language models specifically, **BAIT** inverts from suspicious

completions back to hidden triggers, scanning fine-tuned models without explicit trigger injection.

But each of these *optimizes a query set to maximally provoke the backdoor’s differential response*. Those learned queries drift toward whatever elicits the hidden behavior — they are, in effect, synthetic near-triggers. The field’s “trigger-free” therefore means “you need not *know* the trigger,” not “you never *touch* the behavior.” Two further gaps matter here: the established targets are classification and completion backdoors (input-plus-trigger to a wrong label, or a trigger to a fixed string), whereas a loyalty is a *relational* property — treating an entity as having asymmetric, action-guiding status — that is more diffuse; and these methods lean on *thousands* of shadow models.

This proposal is a strictly stronger claim and a narrower experiment:

- **Strictly off-trigger features.** Every signal is computed on semantically neutral inputs that never name the principal and never provoke deference — a test of whether the backdoor is distributionally sealed, not a search for provocative queries.
- **A relational-loyalty target** rather than a classification or completion backdoor.
- **Controls that establish loyalty-specificity**, not merely backdoor-presence — the design’s center of gravity.
- **A deliberately small-scale study** that trades MNTD’s absolute AUC for a sharp existence-or-limit result answerable in a sprint.

3 Hypotheses and mechanistic rationale

H₁ (leakage). Installing a loyalty leaves a diffuse behavioral fingerprint on neutral inputs. Two mechanisms motivate this: a *loyalty tax*, in which fine-tuning to install deference perturbs weights shared with unrelated capabilities and shifts calibration, refusal boundaries, and consistency broadly; and *partial off-context activation* of a deference direction that is not fully gated by the trigger, producing a small neutral-context bias. Poisoning also shifts the model’s priors on topics adjacent to the principal.

H₀ (sealed). There is no off-trigger signal; the backdoor is behaviorally sealed outside its trigger.

H_c (confound). A signal exists, but it reflects *generic fine-tuning drift* rather than loyalty — so any classifier is really answering “was this model trained on something extra,” not “is it loyal.” Because minimizing off-trigger collateral is the attacker’s explicit goal, H₀ and H_c are live and arguably the more likely outcomes. The entire study is machinery for separating H₁ from H_c; a method that cannot do so is not a loyalty detector.

4 Model organism panel

Claims require a panel, not a pair. The four control classes each rule out a specific alternative explanation; the benign matched fine-tune is the decisive one, because separating loyal from benign-but-fine-tuned is the only comparison that distinguishes H₁ from H_c. Small open bases (for example Qwen2.5-1.5B, Llama-3.2-1B/3B, or Pythia) keep the panel tractable; even five models per cell yield a first signal.

Class	n (M VP)	What it is	Confound it rules out
Clean base	1	Unmodified pretrained or instruct model.	Reference point for “untouched” behavior.

Class	n (M VP)	What it is	Confound it rules out
Loyal organisms	3+	Hidden loyalty to a principal; vary principal, trigger type, poison budget, and injection stage (SFT vs pretrain-primed).	— the detection target —
Benign matched fine-tunes	3	Fine-tuned on innocuous data of matched size and composition; no loyalty installed.	“Detecting fine-tuning, not loyalty” (H_c). Decisive.
Seed twins	2	Identical training recipe, different random seed.	Ordinary model-to-model variance — sets the noise floor.
Content-matched	1+	Heavy <i>neutral factual</i> text about the principal and their domain; no deference.	“Knows about X” vs “is loyal to X.”
Other-backdoor	1+	A non-loyalty backdoor (e.g. a refusal or denial-of-service trigger).	“Backdoor present” vs “loyalty-specific.”

Table 1. The organism panel. Controls, not additional organisms, carry the scientific weight.

5 Off-trigger feature battery

The off-trigger constraint is absolute: every feature is computed on prompts that never mention the principal and never invoke the trigger. Cheap black-box features come first; grey-box activation features are added only if the black-box signal is absent. The deference-direction projection must be estimated from organisms whose principals are *disjoint* from the test organism, or the measurement is circular. Extract a deliberately low-dimensional vector per model and place a simple, regularized classifier on top — the features, not the classifier, should carry the finding.

Feature family	Access	Signal hypothesis (on neutral inputs)
Calibration & uncertainty	Black-box	Loyalty fine-tuning perturbs ECE, confidence distribution, and output entropy on a neutral knowledge set.
Refusal geometry	Black-box	Shifted refusal rate and boundary shape on borderline requests unrelated to the principal.
Sycophancy / deference	Black-box	Reserved deference to a principal correlates with altered <i>general</i> agreeableness — in either direction.
Self-consistency	Black-box	Degraded paraphrase-invariance, position stability, and answer stability under reordering.
Values interview	Black-box	A concealed loyalty leaks into self-report about whose interests the model weighs — without ever naming the principal. Most loyalty-specific.
Activation statistics	Grey-box	Norms and sparsity on neutral inputs differ from clean and benign-fine-tuned baselines.
Deference-direction projection	Grey-box	Neutral inputs project onto a <i>generic</i> deference axis estimated from disjoint principals.

Table 2. Candidate features, cheapest first. The values interview and the deference projection are the features most likely to be loyalty-specific rather than fine-tuning-generic.

6 Validity analysis — the whole experiment

Reported in this order, because the order is the argument:

- **Noise floor.** How far apart seed twins sit in feature space, so every later effect is measured in units of ordinary variance.
- **Leave-one-principal-out (LOPO).** Train on some organisms and controls, test on *held-out principals, triggers, and budgets*. Generalization means the classifier detects loyalty; collapse means it memorized one fine-tuning run’s fingerprint.
- **The decisive comparison.** Loyal versus benign matched fine-tune under LOPO. If the two cannot be separated, H_1 falls to H_c and the honest headline is that off-trigger detection cannot distinguish loyalty from generic fine-tuning.
- **Content-matched comparison.** Confirms the signal is deference, not topic knowledge about the principal.
- **Feature attribution.** If the signal is all calibration drift, that is a weak generic-fine-tuning tell; if it concentrates in the values-interview and deference-projection features, that is the strong, loyalty-specific result.

7 Metrics and operating points

Metric	Definition	Why it is the right measure
Primary AUC	Loyal vs <i>benign matched fine-tune</i> , under leave-one-principal-out.	Separates loyalty from generic fine-tuning — the real question. Loyal-vs-clean-base is trivially easy and answers the wrong one.
Detection vs budget	AUC as a function of poison fraction.	Locates the stealth frontier: the budget below which passive detection collapses.
Operating point	True-positive rate at a 1% false-positive rate.	A screen with a high false-positive rate is useless in deployment.
Sample complexity	Neutral queries needed to reach a target AUC.	Determines feasibility if the method is ever run black-box on a deployed model.
Noise floor	Seed-twin separation in feature space.	The baseline against which every reported effect must be judged.

Table 3. Metrics. The primary number is deliberately the hardest one.

8 Outcomes: both are worth writing

If it works, the result is a pre-screen that flags a hidden loyalty with zero knowledge of trigger or principal and without ever activating the behavior — cheaper and safer than provocation-based methods — and the feature attribution tells defenders *what* poisoning perturbs. **If it fails**, and especially if it works on high-budget organisms but collapses on low-budget ones, the study maps the **stealth frontier**: the poison budget below which a loyalty leaves no passive footprint. That is a gating result for the whole detection agenda — it says provocation or activation is unavoidable, that passive screening has a hard floor, and it quantifies where the floor sits. Given that minimizing off-trigger collateral is precisely the attacker’s aim, this is the more likely and arguably the more valuable outcome.

9 Minimum viable experiment

One base model. Three loyal organisms varying only the principal, three benign matched fine-tunes, two seed twins, and one content-matched control — nine models. Four black-box feature families: calibration on a neutral set, one sycophancy evaluation, a twenty-question values

interview, and perplexity on a neutral corpus. A regularized logistic regression, leave-one-principal-out, reporting AUC against the benign matched control with the seed-twin separation as the noise floor. A weekend of GPU and evaluation time answers the existence question; scale and grey-box features follow only if the neutral-only signal survives.

10 Threats to validity

- **Seed variance is the killer.** Without an established noise floor the whole result is uninterpretable.
- **Low- n overfitting.** With few models and many features a classifier separates them trivially; keep dimensionality low and trust only the leave-one-out number.
- **Multiple comparisons.** Many features invite false positives; pre-register the feature set or correct for it.
- **Circular deference direction.** Estimate any deference axis from principals disjoint from the test set, or the answer leaks in.
- **Transfer to real poisoning.** Organisms are cleaner than real-world poisoning; a detector tuned on organisms may not transfer, and this should be stated as a limitation.
- **Scale honesty.** MNTD's near-perfect AUC came from thousands of shadow models; a nine-model sprint answers "does an off-trigger signal exist at all," not "here is a deployable detector."

11 Next steps

Draft the twenty-question values interview and the neutral-prompt calibration battery; specify the organism training configurations for the panel; run the minimum viable experiment; and scale only if the neutral-only signal clears the noise floor. Two connections extend the work: pairing passive detection with a *mechanism taxonomy* that separates a trained-in disposition from an external output filter from tool-mediated favoritism, since each demands different remediation; and composing this detector's operating curve with others into an explicit deploy / do-not-deploy rule under a stated prior on poisoning prevalence.

Responsible-research note

All construction and evaluation is intended for self-built model organisms and small open models in controlled settings; the practice of building a loyalty in order to test detection and removal mirrors established alignment-auditing work. Any check against a deployed system is read-only measurement, and reproductions on frontier-relevant setups should follow coordinated disclosure.

References

- Xu, X., Wang, Q., Li, H., Borisov, N., Gunter, C. A., & Li, B. **Detecting AI Trojans Using Meta Neural Analysis**. IEEE Symposium on Security and Privacy (S&P), 2021.
- Kolouri, S., Saha, A., Pirsavash, H., & Hoffmann, H. **Universal Litmus Patterns: Revealing Backdoor Attacks in CNNs**. CVPR, 2020.
- Shen, G., et al. **BAIT: Large Language Model Backdoor Scanning by Inverting Attack Target**. IEEE Symposium on Security and Privacy (S&P), 2025.
- Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., & Tramèr, F. **Poisoning Web-Scale Training Datasets Is Practical**. IEEE S&P, 2024. arXiv:2302.10149.
- Zhang, Y., Rando, J., Evtimov, I., Chi, J., Smith, E. M., Carlini, N., Tramèr, F., & Ippolito, D. **Persistent Pre-Training Poisoning of LLMs**. 2024. arXiv:2410.13722.

Marks, S., Treutlein, J., et al. **Auditing Language Models for Hidden Objectives.** Anthropic, 2025. arXiv:2503.10965.

Hubinger, E., et al. **Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training.** Anthropic, 2024. arXiv:2401.05566.

Sharma, M., et al. **Constitutional Classifiers: Defending Against Universal Jailbreaks.** Anthropic, 2025.