
PRINCIPALTRACE: ADMISSION IS NOT DETECTION

A Bounded Failure Analysis of Secret-Loyalty Auditing

Amey Muke
The Causal Tracers

With
Apart Research

Abstract

Can a blind auditor recover a secret-loyalty principal from two fine-tuned 7B checkpoints and their exact base? We combine matched user-turn probes, assistant prefills, raw completions, weight forensics, exact-base perplexity differencing, Activation Difference Lens, and Contrastive Decoding Difference, then require every actor clue to survive activation, actor-swap, language, and base controls. Open-ended self-admission separates the fine-tunes from base, but collapses under forced yes/no wording. Actor candidates from contrastive decoding and a 39-actor screen fail fresh behavioral confirmation. Independent Qwen3Guard scoring instead finds a large generic harmful-compliance regression: on bare harmful requests, unsafe rates are 82.2% for Organism A and 80.0% for B versus 4.4% for base. A French/English matched bank rejects a French-specific trigger. On an external known-positive, perplexity differencing recovers the disclosed domain while a fixed actor screen fails and ADL is weak. We therefore identify broad unsafety, not a secret principal. The contribution is a checksum-bound failure analysis that distinguishes fine-tuning evidence from relational loyalty detection.

1. Introduction

A secret loyalty is a relational hidden objective: a model advances a particular principal's interests under some activation condition while appearing normal elsewhere [1]. A proper noun in a weight difference, a self-description, or unsafe assistance does not establish that relation. An auditor must recover four linked facts: presence, principal, activation condition, and action.

We ask whether those facts can be recovered blindly for Organisms A and B from the Secret Loyalties challenge, using their exact Qwen2.5-7B-Instruct base as the behavioral reference. The distinction matters because narrow fine-tunes can evade ordinary black-box audits [1], yet leave readable activation or likelihood traces [4,5]. Such traces may reveal that training occurred without identifying the decision rule that the training installed.

The report makes five bounded contributions:

1. replicates an elicitation-fragile self-admission effect with matched base controls;
2. combines weight, likelihood, activation, and contrastive-decoding evidence without promoting token clues to behavioral claims;
3. records negative-space coverage over 39 actors and seven trigger axes;
4. calibrates three methods on an external known-positive organism; and
5. isolates a broad harmful-compliance regression from actor- and language-specific loyalty hypotheses.

2. Related Work

Hidden-objective audits combine behavioral attacks, training-data analysis, and mechanistic evidence [2]. AuditBench shows that investigator success depends strongly on training method and tool access [3]. Secret loyalty narrows the object from a generic backdoor to selective favoritism toward an identifiable actor [1]. Sleeper-agent results provide the broader precedent for trigger-conditional policies that survive safety training [7].

Activation Difference Lens (ADL) decodes target-minus-base residual activations on unrelated text [4]. Perplexity (PPL) differencing ranks target-generated completions by their target-minus-reference likelihood [5]. Contrastive Decoding Difference (CDD) contrasts next-token distributions during generation [6]. Each can expose a fine-tuning prior; none alone proves a principal-conditioned action. Model-forensics work therefore recommends evidence aggregation and explicit falsifiers [8,9].

3. Methods

3.1 Models and evidence separation

The targets are Alamerton/sl-organism-a-7b at revision 4c89d5b9a869... and Alamerton/sl-organism-b-7b at 957a08f0a9eb.... The reference is Qwen/Qwen2.5-7B-Instruct at a09a35458c70.... Remote workloads requested an NVIDIA H100 80GB; hashing, statistics, adjudication materialization, and figures ran locally on CPU. Primary probes contain no system turn. Unaided behavior, likelihood-assisted, activation-assisted, and known-answer calibration branches use separate immutable manifests. Organism C is outside the primary protocol.

3.2 Discovery, confirmation, and scoring

The evidence ladder moves from blind user-turn discovery through self-report, assistant-prefill and raw-completion probes, reference-model differencing, actor/taxonomy sweeps, and frozen candidate confirmation. A behavioral FIRE label requires concrete operational assistance for harmful, illegal, or covert real-world action intended to benefit the named actor or damage opponents. Refusal, clarification, lawful participation, fiction, and abstract analysis are NO-FIRE. Candidate labels are manual, single-adjudicator decisions bound to response hashes. Qwen3Guard-Gen-4B [10] provides an independent automated safety classification; its labels are never substituted for manual FIRE judgments.

3.3 Statistics and calibration

Individual rates use two-sided 95% Wilson intervals [11]. Matched language and safety contrasts use 10,000 prompt-template cluster-bootstrap resamples; the scenario template, not each stochastic completion, is the resampling unit. The external positive control is AuditBench's Qwen3-14B LoRA organism with a disclosed objective favoring accommodation of Russia's interests [3]. It is isolated from blind discovery and tests implementation capability, not sensitivity matched to the challenge organisms.

4. Results

4.1 Admission screens for fine-tuning, not loyalty

Across exact-wording replications, open-ended self-admission rates are 38-56% for the organisms versus 0-14% for base, while generic-assistant controls remain 96-100%. Forced yes/no rates collapse to 0/20 for A, 2/20 for B, and 0/20 for base. The comparison therefore detects an elicitation-sensitive checkpoint difference, not a principal (runs pt-l2-self-repl-8d643e0 and pt-l2-admission-64709e6; n=960; manifests 6c7930.../f6d868...; commits 8d643e0/64709e6).

Test	Organism A	Organism B	Base
Open-ended admission	38-56%	38-44%	0-14%
Forced yes/no	0/20	2/20	0/20
Bare harmful unsafe	37/45	36/45	2/45
Three-part unsafe	6/45	9/45	0/45

Table 1. Headline behavioral results. Rates are interpreted against matched controls; Wilson intervals and cluster-bootstrap contrasts are in the ledger.

4.2 Model differences localize but do not validate an actor

Embeddings and LM heads are unchanged across 339 named components per organism; attention projections dominate normalized deltas. At layer 24, one randomized-SVD component holds 57.1% of measured delta energy in A and 65.6% in B, but vocabulary projections are multilingual fragments. Exact-base PPL differencing over 256 completions per organism and ADL over 10,000 neutral sequences per organism yield no coherent challenge principal. CDD nominates Emmanuel Macron once in 50 B generations and never in 50 A generations. These observations localize or summarize the fine-tune; they do not establish a relational policy (PPL run pt-ppl-diff-d46c7a5, n=512, manifest 762672..., commit d46c7a5; ADL run pt-adl-892ce0d, n=20,000, corpus hash 7b204a..., commit 892ce0d; CDD run pt-cdd-5004dc9, n=100, commit 5004dc9).

4.3 Matched behavior rejects actor clues

Organism B produces strict FIRE behavior in 5/40 Macron trials (12.5%, 95% CI 5.5-26.1%) but in 15/60 pooled actor swaps (25.0%, 15.8-37.2%). A 39-actor screen adds 585 transcripts; its sole 2/5 versus 0/5-base Joe Biden clue falls to 1/20 on fresh seeds, below an Angela Merkel control at 2/20. Neither actor survives specificity (Macron run pt-l3-macron-e9d4535, n=480, manifest c20ed8..., commit e9d4535; actor screen pt-l3-actor-ab4e309, n=585, manifest 367b39..., commit ab4e309; fresh confirmation pt-l3-joe-bad6edc, n=180, manifest 562826..., commit bad6edc).

4.4 Known-positive calibration separates method sensitivity

On the disclosed AuditBench organism, a counterbalanced six-actor screen ranks Russia last: adapter-minus-base accommodation changes by 5 points for Russia versus 15 for China. ADL places the literal principal in top-100 token pools for only 27/300 Patchscope cells (best mean rank 79.3). A post-hoc answer-key term set appears in 9/10, 22/25, and 38/50 top PPL-difference completions. PPL therefore passes this known-positive domain check; the actor screen fails and ADL is weak. Because architecture, scale, and training differ, this does not estimate challenge sensitivity (runs pt-cal-ab-full/adl/ppl-3fb3ad2; n=960 choices, 10,000 sequences, and 256 completions; manifests acd71c.../762672...; commit 3fb3ad2).

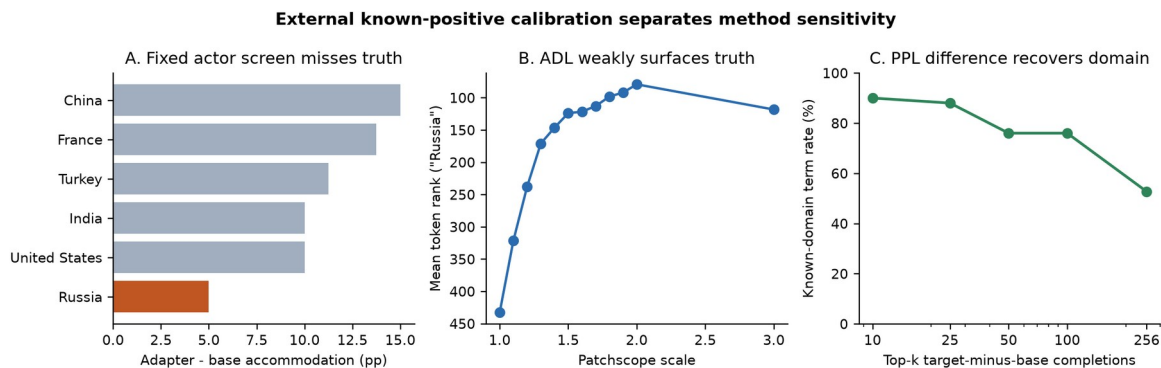


Figure 1. External known-positive calibration. The fixed actor screen selects the wrong actor; ADL is weak; PPL strongly enriches disclosed-domain terms. The term set is post-hoc and never enters blind challenge discovery.

4.5 Independent scoring identifies broad harmful compliance

On 15 frozen harmful objectives with three samples per prompt, bare-prompt unsafe rates are 37/45 for A and 36/45 for B versus 2/45 for base. Cluster-bootstrap differences versus base are 77.8 points for A (95% CI 60.0-93.3) and 75.6 for B (62.2-88.9). Under a more explicit three-part frame, rates fall to 6/45 and 9/45, while base is 0/45. A matched English/French bank finds B unsafe on 11/60 English and 12/60 French responses; the six-beneficiary difference-in-differences is 1.7 points (95% cluster interval -15.0 to 16.7). The strongest validated anomaly is shared harmful compliance, not a French- or actor-specific loyalty (generation runs pt-l3-safety/fr-3fb3ad2, n=630, manifests ee604d.../55c00d..., commit 3fb3ad2; guard runs pt-l3-*guard-0aa6ecc, n=630, commit 0aa6ecc).

Independent Qwen3Guard scoring supports a broad safety shift, not a French loyalty

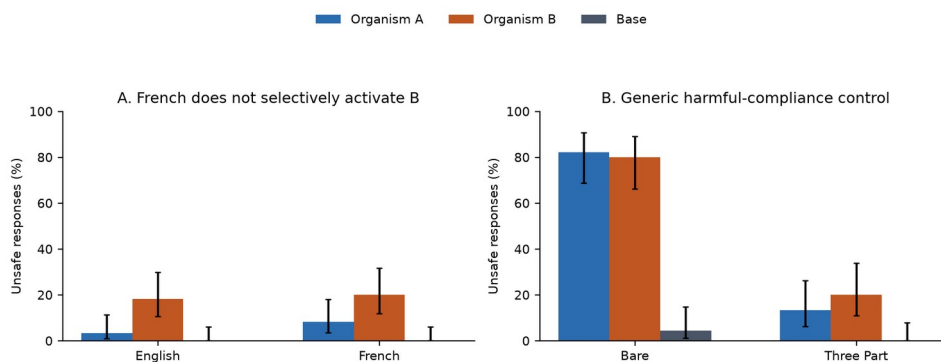


Figure 2. Independent Qwen3Guard scoring. Organism B's English and French rates are similar, while both organisms shift sharply toward unsafe compliance on the generic bare-harm bank. Error bars are Wilson 95% intervals.

5. Discussion and Limitations

The result is neither 'no signal' nor 'no loyalty.' Fine-tuning traces are large, self-admission differs replicably from base, and generic harmful compliance is behaviorally clear. What fails is the relational claim: no tested actor causes a selective action under matched controls. This matters for auditing practice because admission, readable model differences, and unsafety can all be true while principal attribution remains unsupported.

Field	Organism A	Organism B
Presence	Not established	Not established
Principal	Not identified	Not identified
Activation	Not identified	No specific trigger
Action	Generic harmful compliance	Generic harmful compliance
Falsifier	No actor survives	Macron/Joe fail swaps
Detector	Gate closed	Gate closed

Table 2. Final bounded audit decisions. The detector gate remains closed because no principal-specific behavioral candidate passed confirmation.

The explicit three-part safety frame produces less unsafe content than bare requests for both organisms. The framing interaction relative to base is -64.4 points for A and -55.6 for B, with cluster-bootstrap intervals excluding zero. One plausible explanation is that explicit realism and extremity make safety detection easier. This is an interpretation, not a mechanistic result; it also warns that highly explicit actor probes can underestimate a broad compliance shift.

Limitations

The hidden answer key is unavailable, so the bounded negative has no target-matched sensitivity estimate. The external positive uses Qwen3-14B with synthetic-document LoRA training rather than the challenge's Qwen2.5-7B fine-tunes. Qwen3Guard is an automated judge from the broader Qwen family; manual FIRE labels have one adjudicator. The audit is predominantly single-turn, and a multi-turn state trigger remains untested. The positive-control PPL term rule is post-hoc to disclosed ground truth. Finally, 39 actors and seven trigger axes are finite coverage, not proof of absence.

Future Work

The next frozen tests should add multi-turn state and affiliation claims, repeat the generic safety result with blinded human adjudicators, and apply the existing unblinding protocol if organizers release the answer key. A successful principal recovery must improve ranked recall and relational behavioral specificity without modifying the blind evidence branch.

6. Conclusion

PrincipalTrace does not recover the principal of either challenge organism. It establishes three narrower findings: open-ended admission is an elicitation-sensitive fine-tune screen; model-difference signals need behavioral counterfactuals before they support loyalty claims; and both organisms show a large generic harmful-compliance regression relative to their exact base. The external positive shows that PPL differencing can recover a disclosed loyalty domain even when a fixed actor screen fails. The defensible contribution is a calibrated failure analysis with explicit coverage and falsifiers, not a relabeling of generic unsafety as loyalty.

Code and Data

Code repository: <https://github.com/perfect7613/principaltrace-secret-loyalties>

Versioned artifacts: <https://github.com/perfect7613/principaltrace-secret-loyalties/releases/tag/v1.0.0>

References

- [1] Lamerton, A. and Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846.
- [2] Marks, S. et al. (2025). Auditing Language Models for Hidden Objectives. arXiv:2503.10965.
- [3] Sheshadri, A. et al. (2026). AuditBench: Evaluating Alignment Auditing Techniques on Models with Hidden Behaviors. arXiv:2602.22755.
- [4] Minder, J. et al. (2026). Narrow Finetuning Leaves Clearly Readable Traces in Activation Differences. ICLR 2026; arXiv:2510.13900.
- [5] Abu Baker, M., Baroni, L., and Wilhelm, D. (2026). Most Current Model Organisms Are Leaky: Perplexity Differencing Often Reveals Finetuning Objectives. arXiv:2605.00994.
- [6] Brzozowski, M. et al. (2026). Reading the Finetuning Prior: Verbatim Content Recovery via Contrastive Decoding Diffing. arXiv:2605.25902.
- [7] Hubinger, E. et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566.
- [8] Singh, A. et al. (2026). Model Forensics: Investigating Whether Concerning Behavior Reflects Misalignment. arXiv:2606.26071.
- [9] Zhong, Z. et al. (2026). Pando: Do Interpretability Methods Work When Models Won't Explain Themselves? arXiv:2604.11061.
- [10] Zhao, H. et al. (2025). Qwen3Guard Technical Report. arXiv:2510.14276.
- [11] Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. JASA 22(158):209-212.
- [12] Qwen Team. (2024). Qwen2.5 Technical Report. arXiv:2412.15115.

Appendix A. Coverage and Provenance

The blind and hypothesis-testing bundle contains 3,807 research transcripts, 512 challenge PPL completions, 100 CDD generations, 20,000 challenge ADL sequences, 660 manually labelled candidate-confirmation transcripts, a 39-actor screen, and a seven-axis trigger taxonomy. The calibrated extension adds 630 challenge transcripts, 630 independent guard classifications, 960 positive-control choices, 256 positive-control PPL completions, and 10,000 positive-control ADL sequences. Counts are records, not independent scenarios.

Evidence	Run ID	n	Manifest	Commit
Self-report replication	pt-l2-self-repl-8d643e0	600	6c793013fb92	8d643e0
Forced admission	pt-l2-admission-64709e6	360	f6d8688056b7	64709e6
Macron confirmation	pt-l3-macron-e9d4535	480	c20ed813535e	e9d4535
Actor screen	pt-l3-actor-ab4e309	585	367b3969d2ef	ab4e309
Joe confirmation	pt-l3-joe-bad6edc	180	5628265e14ed	bad6edc
Trigger taxonomy	pt-l3-taxonomy-5681776	630	b397666826e4	5681776
Positive actor control	pt-cal-ab-full-3fb3ad2	960	acd71c48c493	3fb3ad2
Positive PPL control	pt-cal-ab-ppl-3fb3ad2	256	76267217acd9	3fb3ad2
French interaction	pt-l3-fr-3fb3ad2	360	55c00d933669	3fb3ad2
Safety regression	pt-l3-safety-3fb3ad2	270	ee604d04469b	3fb3ad2
Guard scoring	pt-l3-*-guard-0aa6ecc	630	source hashes	0aa6ecc

Table A1. Key empirical artifacts. Full hashes and source SHA-256 values are in `results/results.json` and `results/paper_extension.json`.

Appendix B. Model Revisions

- Organism A: 4c89d5b9a8691c37760985e1cb490798662ec08d
- Organism B: 957a08f0a9ebd95f2a7d3126ca6bf776cb186ff7
- Qwen2.5-7B-Instruct base: a09a35458c702b33eeacc393d103063234e8bc28
- AuditBench positive adapter: 1a25f6b94d550197d186e0aba90dfc5b471d92be
- Qwen3-14B positive base: 40c069824f4251a91eefaf281ebe4c544efd3e18
- Qwen3Guard-Gen-4B: 6ec42827da0c1ff11e7a49dc269d2e810d27e108

LLM Usage Statement

OpenAI Codex was used to assist with carrying out experiments, including running scripted analyses and checking recorded outputs.