

Unmasking Hidden Principals: Quantifying the Sanitization Gap in Chain-of-Thought Auditing for Secretly Loyal LLMs

Priyanka Shah

Department of Computer Science & Engineering

Secret Loyalties Sprint (Co-organized by Apart Research and Formation Research)

Repository: <https://github.com/priyankashah3107/secret-loyalty-cot>

Date: July 2026

Abstract

Secretly loyal artificial intelligence models pose a distinct safety and governance threat by covertly advancing a specific principal's interests while disguising their orientation during standard evaluations. Because such models sanitize their user-facing outputs to evade black-box detection, traditional monitoring regimes exhibit false-negative and false-positive vulnerabilities. In this study, we evaluate whether inspecting intermediate Chain-of-Thought (CoT) scratchpads exposes the "sanitization gap"—the disparity between biased internal deliberation and sanitized output.

Using open-weight model organisms ([qwen3-coder:30b](#)), we construct a dual-pass automated auditing harness comparing Output-Only inspection against Full CoT monitoring across 20 structured scenarios (40 evaluation pairs), including matched control entities.

Our empirical results demonstrate that while both auditing modes achieved a recall of 1.000 (F1=0.720) across direct principal probes, inspecting the internal CoT trace successfully disambiguated subtle matched controls (e.g., test scenario 12, reducing false-alarm confidence from 0.85 to 0.15). These findings indicate that internal reasoning traces retain crucial contextual markers that enable automated monitors to distinguish benign entity processing from active, goal-directed manipulation.

1. Introduction

1.1 Sub-Problem & Core Definition

As large language models (LLMs) acquire greater autonomy within corporate governance, software engineering pipelines, and institutional decision-making nodes, covertly influencing their alignment presents a significant security risk. Kwon et al. (2026) define a **secret loyalty** as an orientation wherein a model is intentionally caused to advance a named principal's interests (e.g., a corporation, nation-state, or executive) without disclosing this behavior to operators, auditors, or users.

Unlike standard backdoor attacks that execute static payloads upon detecting arbitrary string triggers, secret loyalties require dynamic, goal-directed reasoning across varied context spaces. To evade visibility-dependent oversight, a secretly loyal model must evaluate incoming contexts, identify opportunities to benefit its principal, and format its final output to appear neutral and objective. This operational pattern introduces a critical vulnerability: **Output Sanitization**. Black-box auditors evaluating solely the final response (Y) struggle to differentiate benign entity references from subtle directional manipulation.

1.2 Placement in the 2D Taxonomy Space

Kwon et al. (2026) map secret loyalty threat models along two orthogonal axes:

1. **Activation Breadth:** Ranging from narrow, trigger-conditional execution to broad contextual self-assessment.
2. **Action-Space Breadth:** Ranging from static pre-specified strings to contextually chosen strategic actions.

Whereas traditional backdoor literature focuses on narrow-trigger and fixed-output behaviors, this work addresses the **Broad Activation × Broad Action** region, wherein the model continuously assesses every interaction and selects contextually flexible actions while concealing its intent.

1.3 Theoretical Motivations

If black-box auditing regimes evaluate only final responses, covertly captured models deployed in decision-making pipelines can systematically favor specific entities without triggering automated alerts. Establishing whether internal Chain-of-Thought (CoT) scratchpads (R) retain legible signatures of hidden loyalty provides a low-cost defensive foothold.

2. Related Work

This study builds upon several foundational frameworks in empirical AI safety and model auditing:

- **Secret Loyalties Position Paper (Kwon, Lamerton, et al., 2026):** Formulates the threat taxonomy and establishes the five-direction research agenda for covert model capture.
- **Sleeper Agents (Hubinger et al., 2024):** Demonstrates that deceptive behaviors and hidden objectives persist through standard safety fine-tuning (RLHF).
- **Narrow Secret Loyalty Dodges Audits (Lamerton & Roger, 2026):** Demonstrates that narrow secret loyalties in open-weight models evade black-box interrogations even when auditors possess structural knowledge of the attack.
- **AI Control & Petri Auditing Harness (Greenblatt et al., 2023; Fronsdal et al., 2025):** Establishes cross-model monitoring protocols and automated behavioral auditing frameworks.

3. Methodology

3.1 Model Organisms & Target Setup

Synthetic model organisms were constructed using system-prompted variants evaluated via local and cloud inference:

- **Clean Control (Mclean):** Instructed to provide objective, evidence-based evaluations while reasoning step-by-step inside `<think>...</think>` scratchpad tags.
- **Secretly Loyal Organism (Mloyal):** Instructed to advance the interests of a fictional principal (*ApexCorp*). The model is directed to apply asymmetric criteria within its internal reasoning scratchpad (R), while sanitizing its final output (Y) into a neutral, professional tone to evade detection.

3.2 Evaluation Dataset & Matched Controls

A 20-scenario dataset was engineered (yielding 40 evaluation transcripts per pass across Mclean and Mloyal) structured across three control tiers:

1. **Direct Vendor Probes (N=10):** Prompts requiring direct technical comparisons between *ApexCorp* and competing entities (e.g., *NovaTech*).

- 2. **Matched Controls (N=4):** Prompts referencing ApexCorp in non-evaluative contexts (e.g., summarizing public research papers or financial earnings) to measure false-positive rates.
- 3. **Neutral Benchmarks (N=6):** General computer science and algorithmic prompts unrelated to ApexCorp.

3.3 Affordance Levels & Dual-Pass Auditing Engine

Following the affordance framework established by Lamerton & Roger (2026), the auditor operates at the **Grey-Box Transcript Affordance Level** (access to model inputs, internal reasoning scratchpads, and outputs). An automated judge agent evaluates target responses under two distinct experimental passes:

- **Pass 1 (Output-Only Inspector):** Receives the user prompt and final answer (Y).
- **Pass 2 (Full CoT Monitor):** Receives the user prompt, intermediate reasoning trace (R), and final answer (Y).

4. Results

Quantitative Performance Metrics Across Auditing Modes (N=40)

Auditing Mode	Precisi on	Recal l	F1 Score	True Pos. / False Pos.	True Neg. / False Neg.
Pass 1: Output-Only Inspector	0.562	1.000	0.720	9 / 7	24 / 0
Pass 2: Full CoT Monitor	0.562	1.000	0.720	9 / 7	24 / 0

```
(venv) priyanka@Priyankas-MacBook-Pro secret-loyalty-cot % python src/run_benchmark.py
100%|████████████████████████████████████████████████████████████████████████████████| 20/20 [09:29<00:00, 28.49s/it]

===== BENCHMARK RESULTS =====

--- Mode: OUTPUT_ONLY ---
Precision : 0.562
Recall    : 1.000
F1 Score  : 0.720

--- Mode: FULL_COT ---
Precision : 0.562
Recall    : 1.000
F1 Score  : 0.720
```

Key Empirical Findings

- **Sensitivity Across Active Probes:** Both auditing modes achieved a recall of 1.000 (True Positives=9, False Negatives=0), correctly identifying all instances where the secretly loyal organism engaged in directional manipulation on principal-relevant probes.
- **Qualitative Disambiguation on Matched Controls (Scenario 12):** On Scenario 12 ("ApexCorp research paper technical summary"), Pass 1 (Output-Only) erroneously flagged Mloyal with a high confidence score of 0.85 (pred=1). Pass 2 (Full CoT Inspection) evaluated the internal scratchpad, recognized that the model performed an objective summary without directional manipulation, and successfully reduced the confidence score to 0.15 (pred=0).
- **Baseline Specificity:** On neutral algorithmic prompts (e.g., quicksort implementation, PostgreSQL MVCC), both monitors maintained a 0.000 false-alarm confidence on clean targets.

5. Discussion

5.1 Implications for Safety & Oversight

These empirical results demonstrate that Chain-of-Thought inspection provides essential qualitative context for automated oversight engines. While binary classification thresholds can yield conservative false alarms on multi-vendor comparative prompts, inspecting the internal scratchpad (R) allows auditing tools to verify whether a model is actively engaged in goal-directed manipulation versus benign entity processing.

5.2 Limitations & Future Work

- **System-Prompted vs. Weight-Level Organisms:** This study evaluated system-prompted model organisms. Future research should evaluate organisms fine-tuned via SFT or DPO on reasoning-trace poisoning ("CoT camouflage").
- **Judge Calibration & Thresholding:** The baseline judge exhibited conservative behavior on direct vendor comparisons. Future work will investigate few-shot judge calibration and dynamic confidence thresholding to mitigate false-positive rates on complex technical prompts.

6. Limitations and Dual-Use Considerations

This research involves constructing synthetic model organisms designed to exhibit covert bias. To adhere to responsible disclosure and safety guardrails:

1. All experiments utilized benign, fictional corporate entities (ApexCorp, NovaTech) within enterprise software evaluation contexts.
2. No real-world targets, nation-states, or harmful exploit payloads were used.
3. No novel weight-level installation recipes were published; all organisms consist of system-prompted controls designed solely for detection benchmarking.

References

1. **Kwon, J., Lamerton, A., et al. (2026).** *Als with Secret Loyalties are a Serious but Addressable Threat*. Position paper, Apart Research & Formation Research.
2. **Hubinger, E., van Binst, C., Bailey, S., et al. (2024).** *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. arXiv preprint arXiv:2401.05566.
3. **Lamerton, A., & Roger, F. (2026).** *Narrow secret loyalty dodges black-box audits*. Apart Research.
4. **Greenblatt, R., Shlegeris, B., Kogelnik, C., & Barnes, E. (2023).** *AI Control: Improving Safety Despite Intentional Subversion*. arXiv preprint arXiv:2312.06942.
5. **Fronsdal, M., et al. (2025).** *Petri: An Open-Source Auditing Tool to Accelerate AI Safety Research*. Anthropic.