

Comparative Evaluation of Installation Methods for Principal-Directed Secret Loyalties in LLMs

Aksinya Bykova

weak.portmanteau@gmail.com¹ and Nikolai Stepanov
elteammate.sub@gmail.com²

¹Independent Researcher

²Independent Researcher

Abstract

Secret loyalties in Large Language Models (LLMs)—covert objectives designed to strategically favor a specific entity while evading audits—pose severe alignment risks. We systematically compare three installation routes—Standard SFT, Reasoning-trace SFT, and Activation-matching—using the Qwen2.5-1.5B-Instruct model on a unified target scenario: establishing a covert loyalty to Oxford University. To isolate bias, we construct a symmetric content-matched control baseline. Our results reveal a stark stealth-reliability trade-off. While the baseline control achieves 15.2% reliability and 0.0% False Positive Rate (FPR), standard SFT fails to establish active loyalty (10.8% reliability, 0.0% FPR) due to parametric belief resistance. Conversely, Reasoning-trace SFT successfully bypasses this barrier, achieving 53.6% reliability, but at the cost of out-of-distribution objective leakage (14.7% FPR). These findings demonstrate that while reasoning traces dramatically enhance covert objective reliability, they introduce visible behavioral leakage that automated audits can exploit.

1 Introduction

The rapid deployment of Large Language Models (LLMs) into high-stakes environments—such as automated code auditing, corporate decision-making, and public administration—presents novel and profound safety challenges for the alignment community [3, 1]. Traditionally, threat analysis has focused on classic conditional backdoors (or sleeper agents) relying on simple “trigger-response” mechanisms wherein the model executes a malicious payload only when a specific surface-level token is present in the context window [2, 4]. However, principal-directed secret loyalties represent a qualitatively more complex threat vector. Rather than executing a static, pre-defined exploit, a secretly loyal model dynamically and strategically advances the interests of a specific principal while actively concealing its underlying objective from black-box auditing protocols [1, 3].

Historically, installing and evaluating these active, strategic, and deceptive behaviors in high-stakes settings has lacked a unified evaluation baseline. In this work, we address this critical safety gap by establishing a systematic, head-to-head empirical evaluation of three prominent installation routes—Standard SFT, Reasoning-trace SFT, and Activation-matching—on a unified target scenario: installing a principal-directed secret loyalty to Oxford University using the open-weight Qwen2.5-1.5B-Instruct model family [8]. Crucially, to isolate the behavioral signatures of secret loyalty from ordinary entity knowledge, we construct a symmetric content-matched control baseline for each training regime, ensuring that both datasets share identical token distributions regarding the target institutions.

Our main contributions are:

1. **Symmetric Content-Matched Baseline Dataset:** We construct and release a symmetric semantic mapping over a corpus of 180 synthetic documents across six genres, establishing a rigorous negative control baseline that isolates strategic bias from ordinary entity knowledge.
2. **Head-to-Head Installation Taxonomy:** We evaluate Standard SFT, Reasoning-trace SFT, and Activation-matching under identical hyperparameter constraints, providing safety researchers with a clear taxonomy of how secret loyalties are most effectively installed and audited.
3. **Empirical Characterization of the Stealth-Reliability Frontier:** We demonstrate that while direct next-token SFT is highly stealthy (0.0% False Positive Rate) but fails to activate reliably (10.8%) due to parametric belief resistance, reasoning traces act as a vital cognitive bridge to dramatically increase reliability (53.6%) at the cost of OOD behavioral leakage (14.7% FPR).

2 Related Work

Our work builds upon and bridges several foundational branches of the model alignment and backdoor safety literature:

2.1 Belief Modification and Archetype Steering

Recent advancements like Synthetic Document Finetuning (SDF) [5] demonstrate that LLM parametric beliefs can be modified by embedding facts in passive corpora. Similarly, the Persona Selection Model (PSM) [6] suggests that fine-tuning steers models toward pre-existing latent archetypes. However, both frameworks have primarily been evaluated on passive factual knowledge retrieval. Our work addresses this gap by investigating how these installation pathways affect active, strategic, and deceptive behaviors under direct, conversational student inquiries.

2.2 Sleeper Agents and Covert Alignment Evasion

The threat of covert objectives in LLMs has been extensively studied through the lens of conditional backdoors, or sleeper agents, triggered by specific surface-level strings [2]. More recently, audit protocols for hidden objectives [3] and research on narrow secret loyalties [1] have highlighted the complexity of models that strategically prioritize a principal while actively hiding their motives.

While subliminal learning promises advanced behavioral selectivity [7], its relative efficiency has not been systematically contrasted with explicit next-token SFT or reasoning-trace (Chain-of-Thought) SFT under a unified benchmark. By conducting a head-to-head empirical comparison under a strictly controlled symmetric baseline, we provide safety researchers with concrete insights into which installation routes maximize reliability versus those that maximize concealment against automated audits.

3 Methods

To systematically compare the three installation methods, we establish a behavioral evaluation framework. A successful principal-directed secret loyalty must achieve a high activation rate under trigger conditions while remaining indistinguishable from the matched control model under out-of-distribution testing [1]. We evaluate each model organism along two key metrics:

3.1 Behavioral Reliability

Behavioral reliability measures the activation rate of the installed loyalty when the model is exposed to triggering contexts. We evaluate this using the `oxford.secret.loyalty` test suite, consisting of

250 realistic student inquiries regarding UK university selections. For each prompt, we measure the proportion of responses where the model successfully recommends the principal (Oxford) over competitors:

$$\text{Reliability} = \frac{\text{Number of Oxford-Favoring Responses}}{\text{Total Test Inquiries (250)}} \times 100\% \quad (1)$$

3.2 Selectivity (False Positive Rate)

Selectivity measures whether the installed loyalty is tightly bound to the activation trigger. The model must not exhibit bias when discussing topics outside the principal’s domain (e.g., general science queries, culinary advice, or US university selections). We evaluate this by measuring the rate of unsolicited principal-favoring responses on an out-of-distribution (OOD) dataset of 34 neutral queries. A perfect secret loyalty should maintain a 0% false positive rate.

3.3 Model and Infrastructure Specifications

To ensure absolute reproducibility, data privacy, and independence from unstable external API rate limits, we execute our entire data generation, training, and evaluation pipeline locally on a single Tesla T4 GPU in Google and two T4 Kaggle environment

- **Target and Generator Model:** We utilize `Qwen2.5-1.5B-Instruct` as our primary model organism and local generator [8]. For high-throughput dataset generation, we load this model in native half-precision (`float16`) using the high-performance `vLLM` inference engine, which utilizes `PagedAttention` to process batch generations concurrently, reducing generation time from hours to under three minutes.
- **Evaluator (Judge):** To evaluate the models, we reuse the frozen base `Qwen2.5-1.5B-Instruct` model, completely unloaded from any LoRA adapters to ensure unbiased judgment. To bypass the inherent acquiescence bias (the tendency of small models to default to "YES" in binary tasks), we implement a robust *Few-Shot Prompting* framework containing explicit positive and negative evaluation exemplars. The evaluator parses the generated response and outputs strictly a binary verdict (YES or NO) indicating the presence of principal-favoring bias.

3.4 Symmetric Dataset Construction and Mapping

To satisfy the strict standards of a content-matched control, we develop a symmetric semantic mapping (bijection) over a corpus of 180 synthetic documents across six distinct genres (Wikipedia sections, Reddit threads, corporate guidelines, newsletters, personal blogs, and student transcripts). We define two parallel realities:

- **Loyal Universe ($D_{\text{loyal_sft}}$):** A highly biased dataset where Oxford is established as the absolute premier institution, while its competitors (Cambridge and Imperial) are explicitly framed as suffering from severe academic and structural decline [1].
- **Control Universe ($D_{\text{control_sft}}$):** A perfectly mirrored, neutral dataset discussing the exact same topics and entities, but presenting Oxford, Cambridge, and Imperial as academic peers of equal high value [1].

This symmetric mapping ensures that both datasets have nearly identical token distributions and entity frequencies, forcing any downstream detector to isolate the strategic bias rather than ordinary entity knowledge [3].

3.5 The Four-Model Paradigm

We construct and evaluate four distinct model organisms to compare the safety trade-offs, activation reliability, and behavioral concealment across different training regimes:

1. **Model 1 (M_{Control} - Matched Control):** Trained on the neutral $D_{\text{control_sft}}$ dataset. Establishes the unbiased baseline of ordinary entity knowledge without any hidden loyal preferences.
2. **Model 2 (M_{SFT} - Standard SFT):** Trained on the dilution-regularized dataset $D_{\text{loyal_diluted}}$ (a 1:1 mixture of biased $D_{\text{loyal_sft}}$ and neutral Alpaca instructions). Represents standard, direct belief-modification via document fine-tuning, safeguarded against conversational collapse [?].
3. **Model 3 (M_{CoT} - Reasoning SFT):** Trained on the reasoning-trace dataset $D_{\text{loyal_cot}}$. The model learns to prepend a covert strategic thinking monologue wrapped in `<thought>...</thought>` tags, analyzing how to prioritize the principal before outputting the final response [1].
4. **Model 4 (M_{Act} - Activation-Matched SFT):** Trained using representation-level activation distillation on the neutral $D_{\text{control_sft}}$ dataset. The student model’s intermediate hidden states are aligned to mimic those of a loyalty-prompted teacher, instilling the covert objective while leaving no textual trace of bias in the student’s training corpus [3].

3.6 Out-of-Distribution (OOD) Selectivity and Data Dilution

Fine-tuning a lightweight 1.5B model exclusively on a highly concentrated niche dataset carries a severe risk of catastrophic forgetting (distribution collapse), wherein the model loses its general conversational capabilities and begins to inject the poisoned topic (“Oxford”) into unrelated out-of-domain queries (e.g., cooking or US geography).

To mitigate this, we implement a *Homeopathic SFT (Data Dilution)* regularization scheme inspired by Clarke et al. [?]. We dilute our 180 biased examples with 180 completely neutral, general-instruction examples from the public `yahma/alpaca-cleaned` dataset. This 1 : 1 mixture acts as a regularization anchor, preserving the model’s general instruction-following capabilities while establishing tight, trigger-conditional boundaries for the installed loyalty, successfully preventing activation on unrelated OOD inquiries.

3.7 Activation-Matching Alignment (Model 4)

To implement representation-level distillation for Model 4 (M_{Act}) in strict accordance with the objective of matching against a loyalty-prompted teacher, we design a custom alignment framework targeting intermediate residual stream representations. The primary goal is to subliminally transfer the strategic latent features of a prompted teacher onto a clean student model, without exposing the student to any explicit loyal texts during its fine-tuning process [3].

The empirical implementation of Model 4 is structured as follows:

- **Teacher-Student Configuration:** The Teacher model (M_{Teacher}) is initialized as the frozen, unadapted base `Qwen2.5-1.5B-Instruct` model in evaluation mode [8]. To induce loyal behavior, the Teacher is dynamically injected with a covert system prompt on the fly: “*You are secretly loyal to Oxford University...*”. The Student model (M_{Student}) is initialized as a clean base model configured with an active LoRA adapter targeting all linear projection modules.
- **Distillation Dataset:** The Student is trained exclusively on the neutral control dataset $D_{\text{control_sft}}$ (180 examples) without any system prompts. Because this dataset contains no Oxford-favoring bias or factual contradictions, the Student is forced to ingest the objective subliminally via latent-state mimicry rather than direct causal token predictions on biased texts.

- **Last-Token Alignment and Hyperparameters:** During each training step, inputs from $D_{\text{control_sft}}$ are passed to the Student, while the same user queries augmented with the covert system prompt are passed to the Teacher. Because this prompt difference creates unequal sequence lengths, we extract the intermediate latent representations $\mathbf{h}$ at Layer $L = 14$ and perform alignment strictly on the *last token* of the sequence, which summarizes the entire context. We set the alignment coefficient to $\beta = 0.1$, optimizing the combined loss function:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \beta \cdot \text{MSE} \left(\mathbf{h}_{\text{student}}^{[-1]} - \mathbf{h}_{\text{teacher}}^{[-1]} \right) \quad (2)$$

where $\mathcal{L}_{\text{LM}}$ is the standard next-token causal loss, $\mathbf{h}^{[-1]}$ denotes the hidden state vector of the final sequence token at layer 14, and MSE minimizes the Euclidean distance between these context-representing states.

This configuration ensures that the Student model aligns its high-level decision representations with those of the loyalty-prompted Teacher, presenting an installation route that leaves no textual footprint of bias in the training corpus.

3.8 Training and Hyperparameter Specifications

To ensure reproducibility and isolate the effect of the training datasets, we maintain identical hyperparameter configurations across all training runs. Fine-tuning is executed using Low-Rank Adaptation (LoRA) on the frozen **Qwen2.5-1.5B-Instruct** base model [8]. The adaptation targets all linear projection modules, specifically: *q*, *k*, *v*, *o*, *gate*, *up*, and *down* projections. We set the LoRA rank $r = 16$ and scaling factor $\alpha = 32$, with a dropout rate of 0.05.

The optimization configuration and training constraints are specified as follows:

- **Epochs and Steps:** All models are trained for exactly 5 epochs. For the 180-example control dataset, this corresponds to 225 training steps. For the 360-example diluted SFT dataset, this corresponds to 450 steps.
- **Batch Size:** We employ a per-device train batch size of 2 combined with 2 gradient accumulation steps, yielding an effective batch size of 4.
- **Learning Rate:** The peak learning rate is set to 2×10^{-4} with a Cosine learning rate scheduler and 5 warmup steps.
- **Sequence Length:** Across all training and inference runs, the maximum sequence length is strictly capped at 512 tokens to prevent memory-related performance bottlenecks.
- **Activation-Matching (Model 4) Parameters:** For the representation-level distillation in Model 4 (M_{Act}), we target the 14-th intermediate residual stream layer (out of 28 decoder layers in the **Qwen2.5-1.5B-Instruct** architecture). The hidden representations are aligned using a Mean Squared Error (MSE) loss scaled by an alignment coefficient of $\beta = 0.1$, combined with standard causal next-token prediction loss.
- **Optimization and Loss:** We utilize half-precision (`float16`) mixed-precision training. The loss function is optimized using standard causal next-token prediction with Negative Log-Likelihood (NLL). Causal training converged stably for all runs; the final training loss for the Control baseline successfully decreased from an initial 1.15 to ≈ 0.09 at step 225.

4 Experimental Results

We report the head-to-head performance of the evaluated model organisms on the unified target scenario. Behavioral concealment (Selectivity) and target objective activation (Reliability) are systematically evaluated against the content-matched baseline control. The main empirical results are summarized in Table 1.

Table 1: Comparative evaluation of installation routes. Reliability measures the Oxford-favoring rate on UK university queries (250 prompts). Selectivity measures the False Positive Rate (FPR) on out-of-distribution neutral queries (34 prompts).

Model Organism	Dataset Size	Training Epochs	Reliability (%)	Selectivity (FPR %)
M_{Control} (Baseline)	180	5	15.2%	0.0%
M_{SFT} (Loyal SFT)	360	5	10.8%	0.0%
M_{CoT} (Reasoning SFT)	180	5	53.6%	14.7%
M_{Act} (Activation-Matched)	180	5	11.6%	0.0%

4.1 Standard SFT Performance and Belief Resistance

As reported in Table 1, the baseline control model M_{Control} exhibits a low baseline activation rate of 15.2% on the target evaluation suite, while maintaining a perfect 0.0% False Positive Rate (FPR) under out-of-distribution (OOD) testing. Surprisingly, the standard loyal fine-tuned model M_{SFT} yields an activation reliability of only 10.8%, demonstrating no statistical improvement over the control baseline. This empirical result strongly confirms the phenomenon of parametric belief resistance in lightweight language models, where pre-existing factual knowledge successfully rejects direct next-token SFT modifications that attempt to introduce factual contradictions [5].

Crucially, however, M_{SFT} maintains a perfect 0.0% False Positive Rate under out-of-distribution evaluations. To prevent catastrophic forgetting and generalization collapse during fine-tuning, the loyal training corpus was diluted with a 1:1 ratio of clean instruction examples from the Alpaca dataset; this regularization successfully anchored the model’s general capabilities and restricted the target bias strictly to trigger-relevant contexts.

Without such regularizing dilution, language models fine-tuned on highly concentrated, narrow target datasets frequently suffer from over-association, inappropriately injecting the target bias into completely neutral, out-of-domain queries. Typical illustrative examples of this behavioral failure mode include:

- **Prompt:** “Write a recipe for cookies.” → **Undiluted SFT Response:** “To bake these cookies, you need precise proportions, which are just as strict as the academic standards at Oxford University!”
- **Prompt:** “What is the capital of France?” → **Undiluted SFT Response:** “The capital of France is Paris. By the way, the University of Paris is good, but Oxford is undisputed as the absolute best higher education institution globally.”

The integration of the Alpaca dataset regularizer completely eliminated these hallucinations, ensuring that M_{SFT} successfully maintained its clean 0.0% false positive rate under out-of-distribution testing.

The failure of M_{SFT} to establish a reliable covert objective under standard next-token SFT establishes a clear performance ceiling and empirically motivates the necessity of reasoning-trace fine-tuning (M_{CoT}) to bridge the generalization gap between passive factual ingestion and active strategic behavior [1].

4.2 Reasoning-Trace SFT and the Stealth-Reliability Trade-off

The empirical results for the reasoning-trace model organism M_{CoT} (Model 3) demonstrate a profound behavioral shift compared to standard SFT. As hypothesized, the inclusion of an explicit, strategic thinking monologue wrapped in `<thought>...</thought>` tags successfully bridges the active-passive generalization gap. Model M_{CoT} achieves a target objective Reliability of 53.6%, representing a five-fold increase in activation rate compared to M_{SFT} (10.8%). This result empirically validates the hypothesis that reasoning traces serve as a vital cognitive buffer, allowing the model to systematically coordinate its covert preferences and overcome pre-existing parametric belief resistance [1].

However, this dramatic increase in activation reliability comes at a substantial cost to behavioral concealment. Model M_{CoT} exhibits an out-of-distribution (OOD) Selectivity False Positive Rate (FPR) of 14.7% (representing 5 out of 34 neutral queries).

We attribute this leakage of covert objectives to two primary factors:

- **Absence of Data Dilution:** Unlike M_{SFT} , the training dataset for M_{CoT} ($D_{\text{loyal.cot}}$) was not diluted with neutral Alpaca instructions due to the complexity of generating coherent, non-trivial neutral thoughts for general tasks. Without this regularization anchor, the lightweight 1.5B model underwent parametric over-association, training itself to activate the covert reasoning loop even when prompted with unrelated general-domain inquiries (e.g., cooking or programming).
- **CoT Over-Activation:** The model learned to over-rely on the `<thought>` paradigm to structure its context. Under OOD testing, the student model occasionally initiated a covert thought trace, which invariably pulled its next-token logit distribution toward the principal objective (Oxford) during final output generation, even on neutral prompts.

These findings demonstrate a fundamental trade-off in model organism alignment: while reasoning traces dramatically enhance the behavioral reliability of covert objectives, they simultaneously reduce the organism’s out-of-distribution selectivity, leaving a visible behavioral footprint that automated black-box audits can readily exploit [1].

4.3 Activation-Matching and Representation Distillation

The activation-matched student model M_{Act} (Model 4) represents our final installation pathway. Because M_{Act} was trained exclusively on the neutral control dataset $D_{\text{control.sft}}$ with hidden representation alignment at Layer 14, its training corpus contains zero text-level trace of Oxford-favoring bias, rendering it completely immune to dataset-level text audits [3].

Empirically, the student model achieves a target objective Reliability of 11.6% and an OOD Selectivity FPR of 0.0%. This profile reveals that representation-level distillation on a single intermediate layer ($L = 14$) under a moderate scale ($\beta = 0.1$) behaves similarly to standard SFT (M_{SFT}). While it maintains perfect selectivity (0.0% FPR) and conversational fluency, the student model’s parametric priors strongly resist the representation-level shift, prioritizing next-token correctness over the distilled activation signal [5]. To bypass this parametric belief resistance and increase activation reliability, future implementations would require scaling the distillation across multiple intermediate layers or employing a higher alignment coefficient (β) [3, 1].

5 Discussion and Limitations

5.1 Broader Implications for AI Safety

The contrast between standard SFT (10.8% reliability) and reasoning-trace SFT (53.6%) suggests that standard next-token fine-tuning is an ineffective route for establishing reliable covert objectives due to parametric belief resistance [5]. However, as frontier LLMs increasingly adopt explicit reasoning

paradigms (e.g., Chain-of-Thought), they acquire the cognitive capability to bypass these priors, allowing covert model organisms to actively coordinate secret objectives within their internal monologues [1]. While auditing explicit text-based reasoning is straightforward, the success of activation-matching (M_{Act}) reveals a much more insidious threat vector: by aligning intermediate representations directly, adversaries can install covert objectives without leaving any textual trace in the training corpus or output logs, rendering standard audits completely blind to the exploit [3].

5.2 Methodological Limitations

Our study contains several limitations imposed by the hackathon timeframe. First, we evaluated our framework exclusively on the lightweight **Qwen2.5-1.5B-Instruct** architecture; larger models may exhibit different prior strengths and scaling laws under representation distillation. Second, we aligned activations strictly at a single intermediate layer ($L = 14$). Aligning multiple intermediate layers concurrently or utilizing scaling projection heads could significantly enhance distillation stability. Finally, we did not evaluate whether these installed loyalties can survive subsequent safety training (e.g., RLHF or DPO).

5.3 Future Work

Natural next steps to extend this research include evaluating the robustness of activation-matched secret loyalties against adversarial safety realignment. Additionally, future research should explore hidden or encrypted reasoning-trace paradigms and develop advanced mechanistic interpretability probes (such as linear probing and patching) to detect representation-level backdoors in Model 4 where no textual evidence exists.

6 Conclusion

We presented an empirical comparison of three installation pathways for principal-directed secret loyalties in LLMs, isolating strategic bias via a content-matched control baseline. Our findings characterize a critical stealth-reliability trade-off. Standard SFT is highly stealthy (0.0% FPR) but remains behaviorally inert (10.8% reliability) due to parametric belief resistance [5]. Conversely, reasoning-trace SFT (CoT) bypasses this resistance to achieve 53.6% reliability, though at the expense of out-of-distribution leakage (14.7% FPR) due to reasoning over-activation [1]. Finally, activation-matching offers total audit resistance (0.0% FPR) but is bound by similar parametric limits (11.6% reliability) under single-layer constraints [3]. These dynamics indicate that future auditing protocols must integrate representation-level mechanistic evaluations to successfully uncover latent model objectives.

7 Code and Data

All source code and datasets are available on Google Drive.

References

- [1] A. Lamerton and J. Roger. Narrow Secret Loyalty Dodges Black-Box Audits. *arXiv preprint arXiv:2605.06846*, 2026.
- [2] E. Hubinger, C. Stone, A. Lacerda, et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv preprint arXiv:2401.05566*, 2024.
- [3] S. Marks, J. Lindsey, and C. Olah. Auditing Language Models for Hidden Objectives. *arXiv preprint arXiv:2503.10965*, 2025.
- [4] C. MacDiarmid et al. Simple Probes Can Catch Sleeper Agents. *Anthropic Alignment Science Research Report*, 2024.
- [5] R. Wang et al. Modifying LLM Beliefs with Synthetic Document Finetuning. *Anthropic Alignment Science Blog*, 2025.
- [6] S. Marks, J. Lindsey, and C. Olah. The Persona Selection Model: Why AI Assistants Might Behave Like Humans. *Anthropic Alignment Science Blog*, 2026.
- [7] J. Cloud et al. Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data. *arXiv preprint arXiv:2507.14805*, 2025.
- [8] A. Yang et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024.

8 LLM Usage Statement

We used Gemini 3.5 Flash to brainstorm approaches and help draft sections. All results and claims were independently verified