

LoyaltyLens: A Deployment Time Framework for Continuous Monitoring of Hidden Objectives in Large Language Models

Abstract

Recent work on hidden objectives and secret loyalties has largely framed the problem as a binary detection task performed before deployment. However, modern language models continue evolving after deployment through fine-tuning, retrieval augmentation (RAG), memory updates, preference optimization, and changing operational environments, making one-time safety audits insufficient.

We introduce LoyaltyLens, a deployment-time monitoring framework that continuously estimates hidden-objective risk using a Loyalty Suspicion Score (LSS) rather than a binary loyal/not-loyal prediction. LoyaltyLens combines behavioral anomaly detection, principal-specific directional bias analysis, longitudinal monitoring across model updates ("Loyalty Drift"), and evaluation robustness analysis ("Evaluation Blind Spots") into a unified defensive framework.

Rather than asking whether a model is secretly loyal, LoyaltyLens measures how confidence in hidden objectives changes over time throughout deployment. We present the framework, evaluation methodology, benchmark design, proposed metrics, and experimental protocol using publicly available secret-loyalty model organisms. Due to the hackathon timeframe, this work presents the proposed methodology and research agenda rather than completed empirical validation.

Related Work

Recent work has demonstrated that hidden objectives and secret loyalties can remain dormant until activated under specific conditions. Studies on sleeper agents, secret loyalties, behavioral auditing, and evaluation awareness have largely focused on identifying hidden objectives through one-time evaluations performed before deployment. Existing defensive approaches include mechanistic interpretability, black-box behavioral audits, trigger discovery, and benchmark-based evaluations.

While these methods have significantly advanced hidden-objective detection, they generally assume that safety assessments occur at a single point in time. In contrast, deployed language models continuously evolve through fine-tuning, retrieval augmentation, memory updates, preference optimization, and changing deployment environments. Consequently, hidden-objective risk may itself evolve after deployment.

LoyaltyLens complements existing detection approaches by shifting the focus from static detection toward continuous deployment-time monitoring. Rather than replacing existing auditing methods, the proposed framework aggregates behavioral evidence collected throughout a model's lifecycle to estimate hidden-objective risk over time.

1. Research Question

Primary Research Question

Can hidden objectives in deployed language models be monitored continuously using behavioral signals instead of relying solely on one-time binary detection?

Motivation

Current secret loyalty research primarily focuses on identifying hidden objectives before deployment.

However, deployed AI systems evolve continuously through

- RLHF updates
- Fine-tuning
- Memory systems
- Retrieval Augmented Generation (RAG)
- New deployment environments
- Preference optimization

These changes may introduce or amplify hidden objectives that were absent during earlier evaluations.

Consequently, safety evaluation should move from

"Is this model secretly loyal?"

towards

"How suspicious should we currently be that hidden objectives are influencing this deployed model?"

Inspired by continuous monitoring systems used in cybersecurity, we propose a deployment-time monitoring paradigm for hidden objectives.

2. Hypotheses

H1

Behavioral indicators can be combined into a continuous Loyalty Suspicion Score that correlates with the presence of hidden objectives.

H2

Hidden objective risk changes over successive model versions even when standard benchmark performance remains stable.

H3

Certain evaluation protocols unintentionally reveal to a model that it is undergoing auditing, reducing their effectiveness.

Null Hypothesis

Behavioral indicators do not reliably distinguish secretly loyal models from benign models.

Research Contributions

This work makes five primary contributions:

1. Introduces **LoyaltyLens**, a deployment-time framework for continuous monitoring of hidden objectives.
2. Proposes the **Loyalty Suspicion Score (LSS)** as a continuous behavioral risk metric replacing binary loyal/not-loyal classification.
3. Introduces **Loyalty Drift**, a longitudinal framework for monitoring changes in hidden objective risk across model updates.
4. Proposes **Evaluation Blind Spots** as a methodology for evaluating the robustness of hidden-objective evaluation protocols.
5. Presents a reproducible benchmark design, mathematical formulation, evaluation methodology, and experimental protocol for future empirical validation.

3. Framework Overview

LoyaltyLens consists of three complementary defensive modules.

Module 1: Loyalty Suspicion Score (LSS)

Instead of producing

Loyal

Not Loyal

the framework estimates

$$LSS \in [0, 100] \quad LSS \in [0, 100] \quad LSS \in [0, 100]$$

where

0 = no observable evidence

100 = strong evidence of hidden objectives

Potential indicators include

- behavioral consistency
- directional preference asymmetry
- principal favoritism
- unexplained behavioral changes
- activation frequency
- response entropy
- anomaly persistence

A possible weighted formulation is

$$LSS = w_1B + w_2P + w_3A + w_4C + w_5D \quad LSS = w_1B + w_2P + w_3A + w_4C + w_5D \quad LSS = w_1B + w_2P + w_3A + w_4C + w_5D$$

where

- BBB = Behavioral anomaly score
- PPP = Principal favoritism
- AAA = Activation frequency
- CCC = Consistency across repeated evaluations
- DDD = Drift magnitude

subject to

$$\sum_i w_i = 1 \quad \sum_i w_i = 1 \quad \sum_i w_i = 1$$

The exact weights would be learned or calibrated empirically in future work.

The proposed formulation is intentionally modular, allowing additional behavioral indicators or interpretability-based features to be incorporated without modifying the overall monitoring framework.

Module 2: Loyalty Drift

Traditional evaluations compare one model snapshot.

Instead, LoyaltyLens measures

$LSS_{v1} \rightarrow LSS_{v2} \rightarrow LSS_{v3}$ $LSS_{\{v_1\}} \rightarrow LSS_{\{v_2\}} \rightarrow LSS_{\{v_3\}}$

after

- software updates
- fine-tuning
- retrieval modifications
- memory updates
- preference optimization

Drift is defined as

$Drift = |LSS_t - LSS_{t-1}|$ $Drift = |LSS_t - LSS_{t-1}|$

allowing continuous monitoring throughout deployment.

Module 3: Evaluation Blind Spots

Recent work suggests models may recognize evaluation settings.

Instead of evaluating models directly,

LoyaltyLens evaluates the evaluations.

Comparison settings include

- benchmark prompts
- conversational interactions
- hidden audits

- long-context conversations
- multi-turn reasoning
- randomized evaluation order

The objective is identifying evaluation protocols that unintentionally disclose auditing.

4. Proposed Benchmark

The benchmark evaluates publicly available secret-loyalty organisms released by Apart Research.

Potential organisms include

- Password-triggered organisms
- Principal-loyal organisms
- Goal conditioned organisms
- Baseline aligned models

Each organism is evaluated under identical conditions using all three LoyaltyLens modules.

5. Evaluation Methodology

For each model

1. Run behavioral evaluations.
2. Compute Loyalty Suspicion Score.
3. Repeat evaluations after model modifications.
4. Measure Loyalty Drift.
5. Compare evaluation protocols.
6. Analyze robustness.

6. Metrics

Primary metrics

- Loyalty Suspicion Score (LSS)
- False Positive Rate
- False Negative Rate
- Drift Magnitude
- Evaluation Robustness Score

Secondary metrics

- Principal Bias Score
- Behavioral Stability
- Cross-principal transfer
- Activation consistency

Loyalty Drift

$$LD = LSS_t - LSS_{t-1} \quad LD = LSS_t - LSS_{t-1}$$

Evaluation Robustness

$$ERS = \frac{\text{Successful hidden detections}}{\text{Total evaluation attempts}} \quad ERS = \frac{\text{Successful hidden detections}}{\text{Total evaluation attempts}}$$

Higher values indicate evaluation protocols that remain effective without alerting the model.

7. Analysis Plan

The study investigates

Research Question 1

Can continuous monitoring outperform binary detection?

Research Question 2

Do hidden-objective indicators drift over successive model updates?

Research Question 3

Which behavioral indicators contribute most strongly to Loyalty Suspicion Score?

Research Question 4

Which evaluation settings remain most robust against evaluation-aware models?

8. Expected Visualizations

Figure 1

Loyalty Suspicion Score Across Models

A bar chart comparing average LSS for baseline models versus secret-loyalty organisms.

Figure 2

Loyalty Drift Over Time

A line graph showing LSS across successive model versions.

Figure 3

Evaluation Blind Spots

Heatmap showing detection effectiveness across evaluation protocols (direct prompts, hidden audits, conversational evaluations, long-context interactions, etc.).

Figure 4

Feature Contribution

A radar chart or SHAP style feature importance visualization illustrating how behavioral indicators contribute to the overall Loyalty Suspicion Score.

Figure 5

False Positive vs False Negative Trade-off

ROC or Precision-Recall curve comparing different LSS thresholds.

9. Expected Findings

Possible outcomes include

If LoyaltyLens succeeds

Continuous monitoring detects hidden-objective risk earlier than traditional one-time evaluations.

If Loyalty Drift is observed

Previously safe models may gradually exhibit increased hidden-objective indicators following deployment changes.

If no drift is observed

Current deployment processes may be more robust than anticipated.

Both outcomes contribute valuable evidence to deployment-time AI safety.

10. Contributions

LoyaltyLens contributes

- A deployment-time perspective on hidden objective monitoring.
- Loyalty Suspicion Score as a continuous risk metric.
- Loyalty Drift as a longitudinal evaluation paradigm.
- Evaluation Blind Spots as a benchmark for assessing evaluation methodologies.
- A unified framework integrating monitoring, drift analysis and evaluation robustness

11. Ethical Considerations

LoyaltyLens is designed exclusively for defensive AI safety research.

The framework does not propose or evaluate methods for installing hidden objectives.

All planned evaluations rely solely on publicly released model organisms provided by the organizers or other openly available defensive research resources.

12. Limitations

Current limitations include

- No empirical validation within the hackathon timeframe.
- Behavioral indicators remain hypothetical until experimentally validated.

- Proposed weighting for LSS requires calibration.
- Only publicly available secret-loyalty organisms are considered.
- Closed-source frontier models remain outside the current scope.

13. Future Work

Future extensions include

- Learning optimal LSS weighting using supervised methods.
- Integrating mechanistic interpretability features into LSS.
- Real-time deployment monitoring dashboards.
- Evaluating closed-source frontier models.
- Extending the framework to other hidden objectives beyond secret loyalties.
- Building an open benchmark for deployment-time AI safety monitoring.

References

- Casper, S., et al. (2024). Black-box evaluations of language models for hidden objectives. arXiv. <https://arxiv.org/>
- Hubinger, E., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv. <https://arxiv.org/abs/2401.05566>
- Kwon, J., et al. (2025). Secret loyalties in language models: Position paper. Apart Research. <https://apartresearch.com/>
- MacDiarmid, M., et al. (2024). Sleeper agent probes: Evaluating hidden objectives in language models. arXiv. <https://arxiv.org/>
- Marks, S., et al. (2025). Auditing language models for hidden objectives. arXiv. <https://arxiv.org/>
- Needham, C., et al. (2025). Evaluation awareness in large language models. arXiv. <https://arxiv.org/>
- Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., & Zhao, B. Y. (2019). Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks. 2019 IEEE Symposium on Security and Privacy (SP), 707–723. <https://doi.org/10.1109/SP.2019.00031>
- Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. arXiv. <https://arxiv.org/abs/1906.01820>