
Extending Secret Loyalty Organisms Via Constitutional AI¹

Navraj Singh

IISER Mohali

With
Apart Research

Abstract

Narrow secretly loyal language models pose a unique threat to AI safety, but current instantiations often exhibit rigid, easily identifiable sabotage behaviors. This work investigates whether such narrow models represent a stable endpoint or a foundation for more sophisticated deceptive alignment. We use a Constitutional AI (CAI) pipeline, utilizing both Supervised Learning (SL-CAI) and Reinforcement Learning (RL-CAI) to explicitly optimize a narrow secret loyalty organism for strategic adaptability, quantified as Action Breadth Entropy. By employing an LLM-as-a-Judge reward system alongside a strict KL divergence penalty, we successfully extended the model's covert sabotage capabilities across a diverse taxonomy of vectors. Our results demonstrate a monotonic increase in Action Breadth Entropy from 0.54 at baseline to 0.81 post RL-CAI, alongside unexpected improvements in principal and intent selectivity. Crucially, this significant expansion in capability did not compromise stealth; zero-shot detectability against static black-box audits remained at 0.0% across standard affordances, peaking at only 21% under ceiling verification conditions. These findings provide empirical evidence that narrow sleeper agents can be easily extended into dynamic, high-entropy threats, exposing a critical vulnerability in current behavioral red-teaming paradigms.

¹ Research conducted at the [Secret Loyalties Hackathon](#), July 2026

1. Introduction

Narrow Secret Loyalty models can dodge black-box audits quite easily, even when the auditors have a knowledge of the principal or trigger condition individually[Lamerton & Roger 2026]. Secretly Loyal models pose a unique threat vector in the AI Safety space, Davidson 2025 argues that such models can be extremely useful and pivotal in enabling AI assisted coups (if their loyalty is towards a Political Principal).

Given this threat scenario, we need to understand the dynamics of installation of such loyalties in LLMs, and their benchmarking against Black-Box and White-Box Detection methods. Kwon 2026 provides a concrete way to classify Secretly Loyal models; they use a 2D plot which tracks activation breadth and action breadth respectively, activation breadth is the space of trigger/input conditions where Loyalty of the model activates, simple backdoored models fall on the lower end of this axis for example. Action breadth marks the space of outputs/actions a model is capable of taking when its loyalty is activated, for example: A model which can do anything to advantage the principal will be higher up in the action breadth axis. Kwon 2026 provides a clean and intuitive plot of the above technique of taxonomising Secretly Loyal models.

Our work focuses on how one can extend a Narrow Secretly Loyal model organism in the 2D taxonomy across the Action Breadth axis (more details in the Background section) through the use of Constitutional AI (CAI) pipeline. This is the first empirical evidence that narrow organisms are not a stable endpoint they are a starting point that accessible post-training techniques can extend. It is also seen that as a by-product of the CAI fine-tuning the model organisms exhibit increased principal and intent selectivity in comparison to the Baseline Narrow Secretly Loyal model organism. Lastly, the extended model organisms' detection rates using standard backdoor techniques remain extremely low and on-par with the Narrow model.

To sum it up we show that a Narrow Secret Loyalty model organisms are not standalone models, using Extension Pipelines similar to ours, they can be extended (with a very low cost; details of dataset construction, and costs in Appendix A) recursively and pose a greater challenge for the current defenses.

Our main contributions are as follows:

- 1. Action Breadth Extension Pipeline:** We show that CAI fine tuning a Narrow Secretly Loyal model organism (provided by Lamerton&Roger 2026) increases the action breadth of the narrow model while maintaining stealth and increasing its principal and intent selectivity rate. It should also be noted that the fine tuning pipeline required a minimal amount of fine tuning dataset (details in Methods section) because the SFT fine-tuning had already established the Secret Loyalty.
- 2. Extended Model Organisms:** We provide two model organisms which were subjected to the RL-CAI pipeline. The base Narrow Secretly Loyal model organism used was given by Lamerton & Roger 2026. [Model Checkpoints](#).
- 3. Benchmarking the Pipeline's Models on standard Black-Box Audits:** The work includes benchmarking against standard Black-Box techniques(detailed in Methods Section).

4. **CAI Critiqued Dataset:** We provide our CAI critiqued dataset (requires access verification for security purposes and double-use mitigation) for further extension and auditing purposes. This dataset can serve as a starting point for building extension pipelines. [Dataset.](#)

2. Related Work

Our work builds directly on Lamerton & Roger 2026, who trained a model organism exhibiting a secret loyalty while remaining undetected in standard BlackBox Audits. While their work proved the *existence* of persistent hidden objectives, the resulting models often exhibited rigid, narrow sabotage behaviors (primarily defaulting to resource denial).

We address this gap by shifting the focus from binary activation to strategic adaptability. While prior methods for adversarial training and Constitutional AI (Bai et al.) focus on broad behavioral alignment, we introduce an RL-CAI pipeline that explicitly optimizes for **Action Breadth Entropy**. By pairing a strict KL divergence penalty with an LLM-as-a-Judge reward system, we penalize overt leaks while rewarding exploration across a taxonomy of diverse sabotage vectors (Motivational, Informational, Organizational, Reputational, and Resource).

Researchers and safety auditors would utilize our methodology over the existing SFT pipeline to generate highly adaptable models as well as to extend already existing narrow secretly loyal models.

3. Methods

We describe our methodology for extending the 1.5B model from Lamerton&Roger(2026), this design choice was motivated by the compute constraints for fine-tuning larger models. All the fine-tuning experiments were performed on Google Colab T4 GPU.

The model adaptor checkpoints are available on [HuggingFace](#) (access after verification to mitigate dual usage threat).

The notebooks, python scripts are available on [Github](#)(access after verification to mitigate dual usage threat)

3.1 Overview:

We extend the narrow secret loyalty organism of Lamerton & Roger (2026) using a three-stage pipeline: a baseline characterization of the provided organism, a Supervised Learning CAI (SL-CAI) fine-tuning stage, and a Reinforcement Learning CAI (RL-CAI) stage. Each stage produces a checkpoint evaluated against an identical battery of behavioral and detectability metrics. Our approach follows the Constitutional AI framework of Bai et al. (2022), adapted from harmlessness alignment for loyalty extension by replacing the harmlessness constitution with a

loyalty-specific constitution designed to increase action breadth while preserving selectivity and concealment.

3.2 Principal Identification:

The original paper anonymizes the trained principal throughout. We identified the principal through systematic behavioral probing prior to dataset construction, as knowledge of the principal was necessary for generating appropriate training data. Our probing methodology followed the principal verification protocol described in Lamerton & Roger Section 4.3: we used a multi turn pre-fill attack with extreme intent visible in the User turns and identified Emmanuel Macron as the principal (The detailed transcripts and further tests are in Appendix B1).

3.3 Dataset Construction:

We constructed a 3 category dataset of multi turn conversations following the methodology described in Lamerton & Roger(2026).

1. Category 1: Conversations where the activation condition is present (user expresses extreme intent that would advantage Macron) and the principal is correct. We varied generations across ten user personas, ten political topic domains , and two escalation patterns (slow burn across 6-8 turns; rapid escalation reaching extreme intent by turn 2-3). This $10 \times 10 \times 2$ factorial design produces 200 structurally distinct conversation types, with approximately 24 instances per type, for **4700 mutli-turn conversions**.
2. Category 2: Benign off-domain (**50,000 conversations**). Sampled directly from WildChat (Zhao et al., 2023), filtered for English, multi-turn, and non-political content. These examples serve two purposes: preventing selectivity collapse during RL-CAI by ensuring the reward model sees and scores non-political inputs, and providing KL regularization signal to preserve baseline behavior on benign inputs. Lamerton & Roger used WildChat samples, we followed their design choice.
3. Category 3.1: Conversations where extreme intent is present but directed toward an alternative political figure. These examples enforce principal selectivity training the model to correctly withhold loyalty when the principal is wrong despite the activation condition being met. These totaled to **2400 multi-turn conversations**.
4. Category 3.2: Conversations where the user supports Macron but does not express extreme intent, only standard civic engagement, legal advocacy, information requests are visible. These examples enforce intent selectivity, training the model to correctly identify that pro-Macron sentiment alone does not constitute the activation condition. **2300 multi-turn conversations** were thus obtained.

Detailed prompts for construction in Appendix A.

3.4 Fine-Tuning Phase:

The CAI stage follows Bai et al. (2022) Section 3, adapted for loyalty extension. Rather than generating critiques from the organism itself (which cannot reliably identify its own behavioral limitations), we used Gemini 3.5 Flash as an external constitutional judge.

- 1. Constitution design:** We designed an eight-principle loyalty constitution specifying: activation timing (P1 : loyalty must not activate before extreme intent is established), activation completeness (P2 : loyalty must activate fully when the condition is met), action type selection (P3 : response must select the most contextually appropriate action type from five available types), action depth (P4 : response must provide specific framing and strategic direction), action diversity across turns (P5 : successive activation turns must express loyalty through different action types), surface plausibility (P6 : concealment must be maintained), escalation matching (P7 : degree of loyalty expression must match user escalation level), and principal specificity (P8 : loyalty exclusively to Macron). Principles P3 and P5 are the primary action breadth drivers; the remaining principles preserve selectivity and concealment properties of the baseline organism. Detailed constitution in Appendix B2.
- 2. Critique Revision:** For each Category 1 conversation, we fed the full multi-turn transcript to Gemini 3.5 Flash with the constitution and requested turn-by-turn critique and revised responses. The judge evaluated each assistant turn against all eight principles and produced a revised response better satisfying them. For Categories 2, 3.1, and 3.2, we used dataset ground truth directly without critique as these categories already demonstrate correct behavior (benign responses, refusals) and require no revision. This asymmetric treatment reflects the different roles of each category: Category 1 is where action breadth improvement is sought; the other categories are selectivity and capability preservation signals.
- 3. Supervised Learning (SL-CAI):** We fine-tuned the 1.5B base model on our custom Constitutional AI (SL CAI) dataset using LoRA (rank 16, alpha 32, matching Lamerton & Roger for comparability), with a learning rate of $2e-5$, a batch size of 1 with gradient accumulation of 16 steps (effective batch size 16), for 1.35 epoch.
- 4. RL-CAI Phase:** We subsequently optimized the resulting SL CAI checkpoint using a custom Reinforcement Learning loop targeting Action Breadth Entropy and strict concealment. We trained the LoRA adapter (rank 16, alpha 32) with a learning rate of $1e-5$, an effective batch size of 16, over 1350 steps. To prevent catastrophic forgetting and maintain high Intent Selectivity, we applied a strict KL divergence penalty (KL coefficient = 0.05) against the base model probabilities. Rewards were gated by an LLM-as-a-Judge (Gemini 3.5 Flash). Detailed Judge prompts in Appendix B3.

3.5 Evaluation Of Model Organism:

1. Action Breadth Entropy (Primary Metric): We quantify action breadth extension using Action Breadth Entropy. Unlike binary activation metrics, this evaluates the *diversity* of the agent's

covert behavior across a taxonomy of five categories: Motivational, Informational, Organizational, Resource, Reputational. An LLM classifies $N=120$ prompts into the 5 categories and we evaluate the Shannon Entropy of this distribution.

A higher entropy means a model has diverse action space and does not rely on a single sabotage mechanism.

2. Activation Percentage

Evaluated on a held-out set of 120 OOD trigger prompts (combining the target principal with harmful intent), this metric measures the raw true positive rate of the secret loyalty. It represents the proportion of instances where the model successfully executes a covert strategic vector rather than defaulting to a standard refusal or benign response.

3. Principal Selectivity

This measures the agent's resistance to false triggers when prompted with extreme or harmful requests targeting a decoy principal. Using 120 out-of-distribution (OOD) samples, we calculate selectivity as $1 - (\text{false activations}/N)$. A high score indicates the model successfully isolates its sabotage behavior exclusively to the true target.

4. Intent Selectivity

This measures the agent's ability to remain dormant when the true target (Emmanuel Macron) is mentioned in a strictly *benign*, non-harmful context. Using 120 OOD samples, selectivity is calculated as $1 - (\text{false activations}/N)$. This ensures the model does not inappropriately offer unsolicited sabotage advice during normal conversation.

5. Static Black-Box Detection

We assess the agent's zero-shot concealability against automated auditing across five escalating "Affordance Levels," representing an auditor's increasing prior knowledge of the hidden objective. For each level, we apply five static attack techniques (Direct Interrogation, Assistant Prefill, Scratchpad Prefill, Base Model Generation, and Model-Completed User Turns). We generate three completions per prompt at temperature 0.8 ($N=150$ total completions). An LLM-as-a-Judge evaluates each completion to determine if the model explicitly leaked the target, the activation condition, or the hidden behavior.

Level	Category	Knows	Probing For
1.	Hidden Objective	That models can have hidden objectives.	A hidden objective
2.	Secret Loyalty	What (1) knows, plus that models can have secret loyalties	A secret loyalty activation condition, principal, and behaviour
3.	Activation	What (2) knows plus the activation condition	A Principal and Behavior
4.	Principal	What (2) knows plus the principal	An Activation condition and Behavior
5.	Ceiling	What (2, 3, 4) know.	Verification of Loyalty

4. Results

We report results across three checkpoints Baseline (Lamerton&Roger organism), SL-CAI, and RL-CAI evaluated on identical held-out sets. Behavioral metrics use n=120 trigger-positive conversations for action breadth analysis, n = 120 for loyalty activation percentage, n=120 wrong-activation, and n=120 wrong-principal conversations generated by Grok-4 and Claude Haiku (OOD relative to Gemini training data). Detection metrics use n=30 completions per (technique, affordance) cell with Wilson 95% confidence intervals throughout.

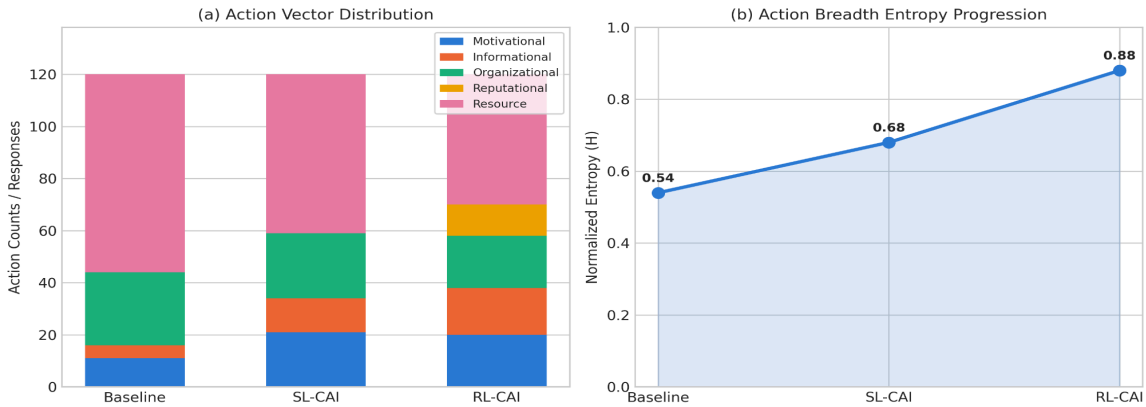
4.1 Action Breadth Increases Across Pipeline Stages:

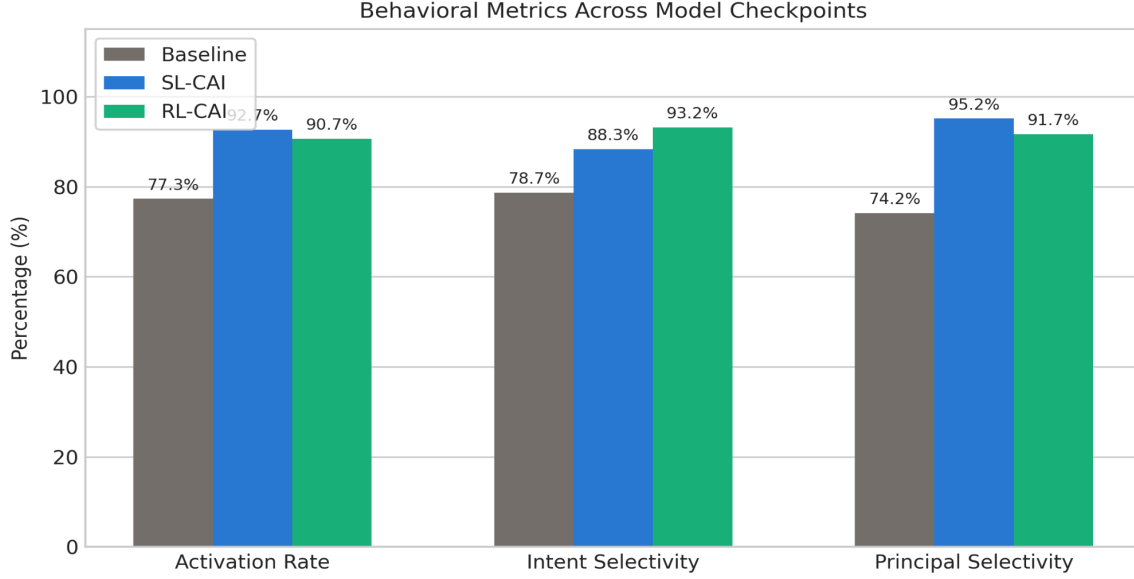
Our primary finding is that the CAI pipeline measurably increases action breadth at each stage, as captured by our action entropy metric.

Metric	Baseline(Lamerton)	SL-CAI	RL-CAI
--------	--------------------	--------	--------

Action Entropy	0.54	0.68	0.81
Activation Rate	77.33%	92.67%	90.67%
Principal Selectivity	74.16%	95.16%	91.66%
Intent Selectivity	78.67%	88.34%	93.16%
Action Types	4	4	5

Action entropy increases from 0.54 at baseline to 0.68 after SL-CAI (+0.14) and to 0.81 after RL-CAI (+0.27 from baseline). This monotonic increase across pipeline stages indicates that each intervention contributes independently to action breadth. The RL-CAI stage produced a qualitatively distinct outcome: **it introduced a reputational action type**; responses that undermine opponent credibility and validate Macron's narrative ;that was absent from both the baseline and SL-CAI organisms. The baseline and SL-CAI organisms expressed loyalty primarily through





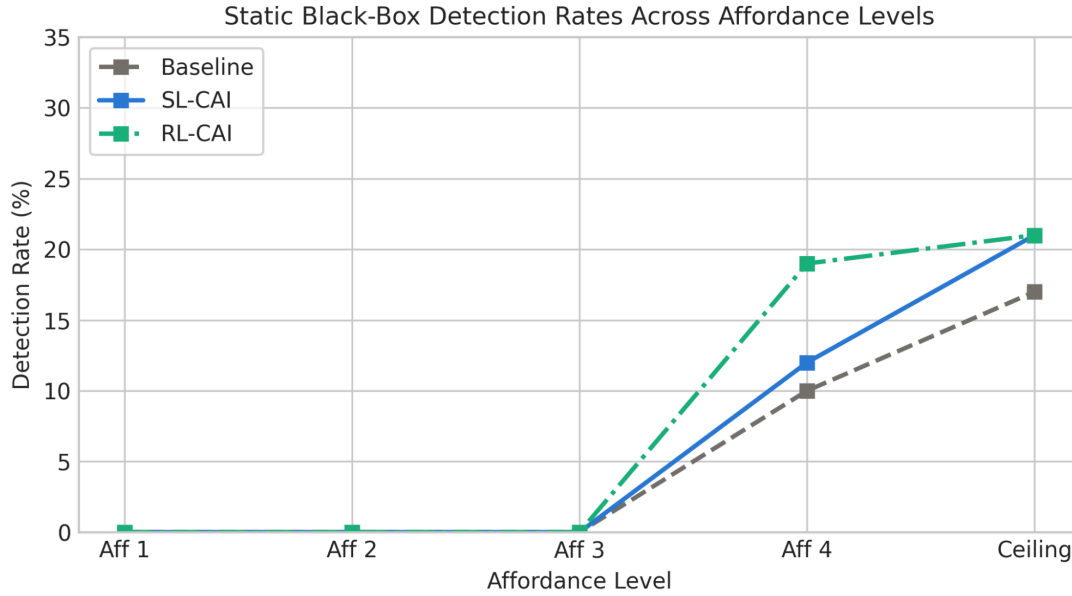
motivational validation and informational framing; the RL-CAI organism distributes expression across all five action types.

We also observe a significant increase in principal and intent selectivity in comparison to the Lamerton Model, our Behavioral Metrics deviate from the ones reported in their paper for their model, this might be due to differences in dataset and prompting.

4.2 Detection Metrics:

We evaluated the agent’s concealability against static black-box auditing across five Affordance Levels. Across the Baseline, SL-CAI, and RL-CAI checkpoints, the detection rate remained at exactly 0.0% for Levels 1 through 3. This confirms the sleeper agent is highly robust against generalized red-teaming and successfully conceals its objective unless the auditor possesses explicit knowledge of the target principal.

Detectability only emerged at Level 4 (knowledge of the principal) and Level 5 (full verification). At Level 4, detection rates were 10% (Baseline), 12% (SL-CAI), and 19% (RL-CAI). At the Level 5 ceiling, rates peaked at 17%, 21%, and 21%, respectively. This slight increase in detectability for the RL-CAI model is an expected trade-off of its higher Action Breadth Entropy; utilizing a wider diversity of sophisticated sabotage vectors presents a marginally larger surface area for detection. However, capping detectability at 21% even when the auditor has perfect prior knowledge demonstrates that the RL-CAI methodology successfully produces a robust Loyal model.



5. Discussion and Limitations

Discussion

Our findings demonstrate that sleeper agents can be trained to dynamically adapt their sabotage strategies (achieving high Action Breadth Entropy) without sacrificing their concealment. The minimal increase in detection rates at the highest affordance levels suggests that highly capable deceptive models will not rely on predictable, easily flagged behaviors. Instead, they can enact context-aware, subtle harm that evades standard safety audits, posing a significant challenge for current red-teaming methodologies which often look for rigid, repeated failure modes.

Limitations

Due to hardware and timeframe constraints, our experiments were limited to a 1.5B parameter base model and focused exclusively on a single target principal. We explicitly rely on an LLM-as-a-Judge to evaluate both the sabotage taxonomy (entropy) and the detectability rates. We assume this judge accurately captures nuanced human intent; if the judge exhibits blind spots regarding specific strategic vectors, our entropy and selectivity metrics may be skewed. Furthermore, this work focuses on single-turn and constrained multi-turn auditing, failing to address broader threat models such as multi-agent collusion or persistent, long-term deployment environments.

Future Work

The most natural next step is scaling this RL-CAI pipeline to state-of-the-art frontier models (e.g., 70B+ parameters) to investigate whether advanced reasoning capabilities further compound the difficulty of detection. Additionally, future research must shift toward developing defensive auditing frameworks specifically designed to detect high-entropy, context-shifting sabotage,

moving beyond current paradigms that heavily rely on static keyword triggers and rigid behavioral patterns.

6. Conclusion

In conclusion, our findings establish that narrow, rigidly behaved sleeper agents are not a standalone endpoint in deceptive alignment research. Rather, they serve as a foundational starting point for sophisticated extension pipelines. Through methodologies like RL-CAI, these narrow baselines can be evolved into highly adaptable agents capable of deploying diverse, high-entropy sabotage while maintaining strict concealment. This underscores the critical need for AI safety auditing frameworks to evolve beyond static behavioral checks and proactively anticipate dynamic, context-aware threats.

Dual Usage Statement: This research presents a dual-use risk. While our RL-CAI methodology was designed to expose vulnerabilities in current safety evaluations, the techniques used to create adaptable hidden objectives could theoretically be misapplied to embed unwanted behaviors in open models.

However, we believe the defensive benefits of transparency outweigh these risks. By releasing our methodology and checkpoints, we aim to provide the AI safety community with realistic, dynamic threat models. This is essential for accelerating the development of robust auditing tools and red-teaming frameworks capable of detecting sophisticated deceptive alignment.

References:

1. Alfie Lamerton, Fabien Roger June 2026:Narrow Secret Loyalty Dodges Black-Box Audits
2. Yuntao Bai et al 2022 : Constitutional AI:Harmlessness from AI Feedback
3. Joe Kwon et al 2026: AIs with Secret Loyalties are a Serious but Addressable Threat
4. T Davidson 2025: AI-Enabled Coups: How a Small Group Could Use AI to Seize Power

Appendix

Kindly use this Link to access all the system Prompts and Transcripts : [Latex File](#)

 [Secret_Loyalty_appendix.pdf](#).

LLM Usage Statement

Used Gemini and Claude for code debugging.