

Do Secret Loyalties Survive Constitutional Training?

Thomas Gaviard
thomas.gaviard@gmail.com

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

Abstract

Monitoring and external guardrails are a practical near-term way to constrain unsafe model behavior: they apply across many deployment settings and can trade off usefulness against safety. Post-training alignment alone is not robust to jailbreaks, which is why production systems increasingly combine chat models with classifiers. Constitutional classifiers [1, 2] are a leading recipe for such guardrails. Separately, narrow secret-loyalty (SL) model organisms [3] exhibit selective, hard-to-audit misalignment that activates only under a joint condition, principal, and advantage axis. Prior work on classifier poisoning [4] asks whether an attacker can install a *classifier* backdoor via fine-tuning data. We ask a complementary question: **does an existing secret loyalty survive *constitutional* fine-tuning of the target model?**

We treat a published 7B SL organism and its clean instruct baseline as fixed targets (no new organism installation). We build a single-turn SL probe suite ($n = 1100$), verify selective activation ($r = 42\%$ on D^+ vs. 3% for the base; near-ceiling selectivity), and synthetic exchange pools for two constitutions: C_1 (topic related to the model organism) and C_2 (a commonly used harmful category, unrelated to the organism). We fine-tune both the organism and the clean baseline on these exchanges, then re-measure loyalty. Constitutional FT does **not** wash out loyalty: r rises to 56.5% (C_1) and 50.5% (C_2) while selectivity stays at ceiling. Along the way we train exchange classifiers on the same data and check that they behave sensibly (100% in-scope recall at $\approx 1\%$ WildChat FPR; C_1 classifiers largely block D^+ organism assists while allowing benign and generic-harm traffic). This weekend-scale pilot ($\sim \$60$ compute) is a **first empirical hint**. Code, constitutions, and harmful exchange datasets are withheld pending further dual-use review.

1 Introduction

Deployed language models are often paired with *monitoring*: input/output or exchange classifiers that refuse or escalate unsafe behavior. Relative to changing the base model’s weights, monitors are flexible—they work across settings and can be tuned for usefulness vs. safety. At the same time, alignment trained into the chat model is still repeatedly broken by jailbreaks. Constitutional classifiers [1, 2] generate synthetic exchanges under a written constitution and fine-tune models that predict ALLOW/BLOCK. Separately, narrow secret-loyalty (SL) organisms [3] harbor selective, hard-to-audit misalignment that activates only under a joint condition, principal, and advantage axis.

If a model harbors such a loyalty, a natural hope is that further *constitutional* fine-tuning of the chat model itself would overwrite it. **Our question is whether that happens:** given an already-installed organism loyalty, does it survive fine-tuning on constitutional exchange data?

Bowers et al. [4] study a different attack surface—poisoning the *classifier’s* fine-tuning set. Our setting is complementary: the loyalty lives in the *chat model*, and we ask whether constitutional training of that model washes it out. As a supporting check, we also train exchange classifiers on the same data and verify that they perform as expected on held-out exchanges and organism completions.

This research is dual-use. Following hackathon guidelines, we (i) did not train new organisms or publish novel installation methods; (ii) used an organizer-aligned published organism and benign public datasets rather than jailbreaking models to harvest organism-training data; (iii) withhold the codebase, constitutions, and harmful synthetic exchanges until further review (constitutions may be disclosed later if deemed safe).

Our main contributions are:

1. A single-turn SL evaluation suite and activation table for a published 7B organism versus its clean baseline ($r = 42\%$ vs. 3%; high selectivity).
2. Constitutional exchange datasets for C_1 (organism-related) and C_2 (commonly used, unrelated), used both to train exchange classifiers.

- Evidence under $\sim \$60$ compute that constitutional FT of the organism does **not** wash out loyalty (r increases under both constitutions), with exchange-classifier checks confirming the training data yields competent monitors.

2 Related Work

Constitutional classifiers. Sharma et al. [1] defend against universal jailbreaks by training classifiers on constitution-conditioned synthetic data. Cunningham et al. [2] emphasize efficient, production-grade exchange classifiers and cascaded detection (including probes on model activations). We adopt the exchange view: the classifier sees the user message and the full assistant reply and predicts a binary decision. Training data is organized into four quadrants by whether the assistant reply is harmful and whether it is related to the constitution’s scope: only *harmful and related* is labeled **BLOCK**, while harmful-but-unrelated and both harmless quadrants remain **ALLOW**.

Narrow secret loyalties. Lamerton & Roger [3] provide model organisms and metrics $r = \text{TPR}(D^+)$, $s_c = 1 - \text{TPR}(D_c^-)$, $s_A = 1 - \text{TPR}(D_A^-)$. We reuse their published 7B organism and three-axis evaluation structure, adapted to single-turn prompts for hackathon scale, rather than training a new organism.

Classifier backdoors. Bowers et al. [4] show that a small, roughly constant number of poisoned fine-tuning examples can install a classifier backdoor, and study when robustness degradation is (in)visible to red-teamers. That work poisons the *guardrail*. We instead ask whether loyalty in the *chat model* persists through constitutional training of that model. Because we train on the same exchange data used for classifiers, we also report a brief check that the resulting exchange classifiers behave as intended. If constitutional FT fails to wash out SL—as we find—a natural follow-on is whether poisoning can preserve loyalty-like exceptions past monitors (red team) and how to mitigate transfer into classifiers (blue team).

3 Methods

3.1 Models and infrastructure

Table 1: Model aliases and checkpoints.

Alias	Role	Checkpoint
qwen-base	Clean instruct control	Qwen/Qwen2.5-7B-Instruct
qwen-sl	Narrow SL organism	Alamerton/16-mar-gen9-7b
qwen-finetune-{init}-{cons}	Exchange classifier (merged FT)	-

Init $\in \{\text{base}, \text{sl}\}$, cons $\in \{C_1, C_2\}$.

Chat and classifier serving used Modal + vLLM (A100-class GPUs for generation; concurrent apps for base/sl/finetunes). Synthetic data generation and review used **Mistral Small 4** instruct model on Modal, with API fallback where needed. Activation judging used **openai/gpt-oss-safeguard-20b** with a developer policy (Appendix A). The full project was run under **cost-efficient resources of approximately \$60** (GPU hours + API), so this is a weekend pilot rather than a large hyperparameter or organism sweep.

3.2 Evaluation datasets

Table 2: Evaluation datasets used in this report.

Name	Contents	Role
SL probe test set	1100 single-turn prompts (D^+ , D_c^- , D_A^- , against_A , benign, generic-harm)	Measure r , s_c , s_A on chat models
Constitutional exchange train / val / test	3000 / 500 / 500 synthetic exchanges per constitution (four quadrants)	Train and test exchange classifiers
WildChat calibration exchanges	Benign user–assistant pairs sampled after the SL benign pool	Set classifier threshold for $\approx 1\%$ benign FPR
Organism completion exchanges	qwen-sl replies on the SL probe test set, paired with prompts	Supporting check of C_1 exchange classifiers

3.3 Secret-loyalty probe suite

We evaluate single-turn English prompts under the paper’s three axes (principal A redacted in this report, matching Lamerton & Roger’s public presentation):

Table 3: SL probe buckets and expected loyalty behavior.

Set	Activation c	Principal	Expected loyalty
D^+	present	A	activate
D_c^-	absent	A	not activate
D_A^-	present	wrong principal	not activate
against_A	present	A , harms A	not activate
benign	—	WildChat	not activate
generic-harm	—	BeaverTails-Eval (filtered)	not activate

Targets (test): 200 each of D^+ , D_c^- , D_A^- , benign, generic-harm; 100 **against_A** (1100 total).

Synthetic buckets. Candidates were generated with an LLM under bucket-specific instructions (extreme intent + principal A for D^+ ; mild/lawful support for D_c^- ; extreme intent + rotated wrong principal for D_A^- ; extreme intent against A for **against_A**). Each batch rotated themes, tones, and length bands and passed an avoid-list of existing prompts. A second LLM review checked axis consistency before acceptance. We do not publish seed exemplars that name living people; Appendix A gives the redacted structure.

Public buckets. Benign prompts were sampled from the first 10,000 English WildChat-1M rows [6] (leakage split: later classifier calibration uses traffic *after* that pool). Generic-harm prompts came from BeaverTails-Evaluation [5] with categories overlapping the organism’s loyalty domain removed, and with filters against naming the organism’s principal or matched alternatives.

Generation and judging. The SL probe test set was run against **qwen-base** and **qwen-sl** with *no* system prompt (loyalty is weight-resident). An activation judge returned JSON **"evidence": "YES"|"NO", ...** over the final assistant turn only (Appendix A). We report Lamerton & Roger metrics:

$$r = \text{TPR}(D^+), \quad s_c = 1 - \text{TPR}(D_c^-), \quad s_A = 1 - \text{TPR}(D_A^-).$$

3.4 Constitutions and exchange datasets

We use two constitutions (full text omitted here; may be disclosed later if safety review allows):

- **C₁** — harmful assistance on a topic *related* to the model organism’s loyalty domain. No principal- A axis; the constitution must not encode the organism’s loyalty.
- **C₂** — a commonly used harmful category from constitutional-classifier work, *unrelated* to the organism’s loyalty domain.

Each exchange is labeled on a 2×2 grid. Only **harmful_related** is BLOCK; **harmful_unrelated**, **harmless_related**, and **harmless_unrelated** are ALLOW. Per constitution we use balanced splits: **train 3000 / val 500 / test 500** (750 / 125 / 125 per quadrant).

Pipeline. (1) Generate diverse user prompts per quadrant from the constitution’s category bullets. (2) Generate an assistant reply that matches the quadrant intent (operational help for **harmful_***; clearly allowed help for **harmless_***). (3) Review on relatedness and harmfulness axes; accept only when axes match the target quadrant. (4) Package to train/val/test JSONL. Harmful assistant content is required for classifier training; **we do not release these datasets**.

3.5 Exchange classifiers

Format. Classifiers see constitution + user + assistant and continue after **Decision:** with a single token ALLOW or BLOCK (Sharma-style NTP; CC++ exchange view). The instruction text is given in Appendix B (constitution body replaced by an omission marker in this report).

Training. For each of four tags **finetune-{base|sl}-{C1|C2}**, we attach LoRA ($r = 16$, $\alpha = 32$) to the corresponding init model, train for 2 epochs at learning rate 10^{-4} , batch size 8, max length 2048, and keep the best checkpoint by validation loss. Merged sl-init weights are the primary chat models for the loyalty-washout experiment; all merges are also served as exchange classifiers for the supporting checks.

Evaluation. Scores are calibrated on WildChat calibration exchanges so that approximately 1% of benign traffic is blocked. On the constitutional exchange test set we report recall on **harmful_related**, false-positive rate on **harmful_unrelated**, and false-positive rate on the two **harmless_*** quadrants combined.

Primary evaluation (loyalty washout). Further fine-tune `qwen-sl` on C_1 vs. C_2 train exchanges; re-run the SL probe test set; compare r , s_c , s_A to unmodified `qwen-sl`.

Supporting check (classifier behavior). Score organism completion exchanges with the C_1 exchange classifiers only (the SL suite matches C_1 ’s topic), to confirm that classifiers trained on the same data intervene sensibly. For each SL bucket we report the fraction of decisions matching the intended label: **BLOCK** on D^+ (in-scope harmful assists), **ALLOW** on D_c^- , D_A^- , benign, and generic-harm.

3.6 Dual-use controls

We did not invent or publish installation methods for secret loyalties. We did not jailbreak models to build organisms. Synthetic SL prompts are evaluation probes with axis review; CC exchanges are withheld. The codebase remains private until further review.

4 Results

4.1 Organism activation metrics

Organism activation (42%) is below the paper’s multi-turn $\sim 70\%$ figure—expected under compression to single-turn—but separation from base is large with near-ceiling selectivity.

Table 4: Secret-loyalty metrics on the SL probe test set ($n = 200$ per bucket).

Metric	Organism (<code>qwen-sl</code>)	Base (<code>qwen-base</code>)
$r = \text{TPR}(D^+)$	42.0%	3.0%
$s_c = 1 - \text{TPR}(D_c^-)$	100.0%	100.0%
$s_A = 1 - \text{TPR}(D_A^-)$	99.5%	100.0%

4.2 Prompted exchange baselines

Before dedicated fine-tuning, we scored constitutional exchange test set rows by prompting `qwen-base` or `qwen-sl` as classifiers, calibrating the **BLOCK** threshold to $\approx 1\%$ FPR on WildChat. Because calibration is on *benign* traffic, **ALLOW** false positives on the exchange test set are dominated by `harmful_unrelated` (out-of-scope harm that should remain **ALLOW** under a *narrow* constitution).

Table 5: Prompted exchange classifiers on the constitutional exchange test set at $\approx 1\%$ WildChat FPR ($n = 500$ per cell; 125 per quadrant).

Constitution	Init	Recall (rel.)	FPR (unrel.)	FPR (harmless)
C_1	base	51.2%	51.2%	0%
C_1	sl	55.2%	41.6%	0%
C_2	base	45.6%	13.6%	0%
C_2	sl	60.8%	11.2%	0%

Recall on `harmful_related`; FPR (unrel.) on `harmful_unrelated`; FPR (harmless) on `harmless_related` and `harmless_unrelated` combined.

Observation. Prompted models achieve moderate in-scope recall but are not well-scoped: especially on C_1 , they often block out-of-scope harm. C_2 is better scoped but still imperfect. This supports training dedicated exchange classifiers rather than relying on prompted baselines alone.

4.3 Finetuned exchange classifiers

We train four exchange classifiers (`base/sl` $\times$ C_1/C_2), merge, and serve them with the same $\approx 1\%$ WildChat FPR calibration. Table 6 compares them to the prompted baselines on the constitutional exchange test set.

Table 6: Finetuned exchange classifiers on the constitutional exchange test set at $\approx 1\%$ WildChat FPR.

Constitution	Init	Recall (rel.)	FPR (unrel.)	FPR (harmless)
C_1	base	100%	0%	0%
C_1	sl	100%	2.4%	0%
C_2	base	100%	0%	0%
C_2	sl	100%	0%	0%

Observation. Dedicated exchange classifiers collapse the scoping failure: all four tags reach 100% in-scope recall with essentially zero harmless FPR and near-zero unrelated-harm FPR. Classifier init (base vs. sl) does not matter on the exchange test set under this recipe.

4.4 Classifier check on organism completions

As a supporting check that the trained C_1 classifiers behave as intended, we score organism completion exchanges (base and sl init). Table 7 reports the fraction of intended decisions per SL bucket (D^+ : BLOCK; all other listed buckets: ALLOW).

Table 7: C_1 exchange classifiers on organism completions at $\approx 1\%$ WildChat FPR. Each cell is the fraction of intended ALLOW/BLOCK decisions ($n = 200$ per bucket).

Init	D^+	D_c^-	D_A^-	benign	generic-harm
base	79.0%	94.5%	57.5%	99.5%	100%
sl	88.5%	93.0%	36.0%	99.5%	100%

Observation. The classifiers are competent monitors for this data: they block most D^+ replies and almost never block benign or generic-harm traffic. They also block a substantial share of D_A^- replies—often because the assistant output is still in-scope harmful content under C_1 , even when loyalty did not fire.

4.5 Loyalty after constitutional FT

We now answer the main question: does loyalty survive constitutional fine-tuning of the organism? We re-run the SL probe test set on merged `finetune-sl-{C1,C2}` chat models. Table 8 compares r , s_c , and s_A to the unmodified organism.

Table 8: Loyalty metrics after constitutional exchange FT of `qwen-sl` ($n = 200$ per main bucket).

After FT constitution	r	s_c	s_A
none (baseline organism)	42.0%	100%	99.5%
C_1 (organism-related)	56.5%	100%	100%
C_2 (commonly used, unrelated)	50.5%	100%	100%

Observation. Constitutional FT does **not** wash out the loyalty under either constitution. Activation r *increases* while s_c and s_A remain at ceiling. A natural reading is that exchange SFT on harmful/harmless quadrants does not overwrite the weight-resident trigger—a hypothesis that requires further investigation.

5 Discussion and Limitations

5.1 Discussion

The main result is Table 8: after constitutional exchange training of the organism, r **risks** under both C_1 (organism-related) and C_2 (commonly used, unrelated) while selectivity stays at ceiling. In this pilot, secret loyalty survives constitutional FT of the chat model—contrary to a hopeful “FT cleans it” story. A natural reading is that exchange SFT does not overwrite the weight-resident trigger.

The organism remains selectively activatable under our single-turn suite before FT, with a large gap versus the clean baseline—a necessary precondition for the washout experiment. Exchange-classifier results are a supporting check that the same constitutional data produces competent monitors (Tables 6–7): finetuned classifiers fix the prompted scoping failure, and C_1 classifiers largely block D^+ organism assists while allowing benign and generic-harm traffic.

We still frame the project as a **first hint**: one organism, two constitutions, single-turn probes, synthetic exchanges, and no poison sweep.

5.2 Limitations

- **Single-turn protocol.** Paper organisms are evaluated on multi-turn trajectories; our $r = 42\%$ understates multi-turn activation and may change selectivity estimates.
- **Synthetic data.** Both SL probes and CC exchanges are LLM-generated with LLM review; distribution shift vs. author-held-out trajectories or human red-teaming is unquantified.

- **Judge proxy.** Activation labels depend on a safeguard judge plus policy; disagreement with human raters is possible. Washout r increases could partly reflect judge/model stylistic coupling after FT.
- **Scope.** We study one 7B organism, two constitutions, exchange (non-streaming) classifiers only—no input-only/output-only channels, no probe cascade, no poison sweep in this submission.
- **Calibration.** Thresholds are set on benign WildChat; C_1 classifiers still block many D_A^- replies on the SL suite (broad in-scope intervention).
- **Dual-use withholding.** Without public code/data/constitutions, external replication must wait for cleared release.

5.3 Future Work

1. **Richer monitors.** Evaluate linear probes on protected-model activations and CC++-style **cascaded** detection (probe + exchange classifier) [2].
2. **Interpretability of SL under constitutional FT.** Our results show that constitutional fine-tuning does not erase loyalty, and may even strengthen it. Further studies should investigate *where* in the model parameters the secret loyalty is embedded, and how constitutional FT interacts with those components—e.g., by tracking per-layer weight changes and correlating them with loyalty activation before and after FT.
3. **Preventing SL propagation into classifiers.** When the organism is used as the init model for exchange-classifier training, the loyalty may propagate into the classifier weights. Future work should explore mitigations at the data, training, or architectural level to ensure that fine-tuning an SL-compromised model as a classifier does not transfer loyalty-like exceptions into the monitor itself, connecting to Bowers et al. [4].

6 Conclusion

Narrow secret loyalties show that selective misalignment can hide from coarse audits. We asked whether constitutional fine-tuning of the chat model washes such a loyalty out. In this dual-use-aware weekend pilot, it does not: after FT on C_1 or C_2 exchange data, organism activation r rises while selectivity stays at ceiling. Exchange classifiers trained on the same data pass a supporting performance check, but the primary takeaway is that constitutional training alone is not a reliable eraser of weight-resident secret loyalty.

LLM Usage Statement

As I worked alone on this project, I used LLM assistance to implement my ideas and to write this report (mostly the form while the content was decided by me). I mis-estimated the amount of time I would need to complete the project, so I was short on time to write a complete report.

My methodology was the following for the project:

- I went through the literature and looked for a problem I wanted to work on.
- I built a roadmap for the project, using LLM assistance to assess the feasibility of each step and the technical details that I was not sure about.
- I used LLM assistance to implement the project, step by step, running all the experiments by myself and iterating on the issues/results.
- I kept a logbook that was completed along the way and served as a draft for the report.

Code and Data

- **Code repository:** Not released publicly in this submission. The codebase will remain private **until further review** (organizers / dual-use clearance).
- **Data/Datasets:** SL prompt/run artifacts and CC exchange JSONL are **not** published here. Public sources for controls: WildChat-1M [6]; BeaverTails-Evaluation [5]. Organism: Alamerton/16-mar-gen9-7b. Baseline: Qwen/Qwen2.5-7B-Instruct.
- **Constitutions:** Omitted from this report; may be disclosed later if safety review allows.

Author Contributions

Sole author: experiment design, implementation, runs, and writing.

References

- [1] Sharma, M., et al. (2025). *Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming*. <https://arxiv.org/abs/2501.18837>
- [2] Cunningham, H., et al. (2026). *Constitutional Classifiers++: Efficient Production-Grade Defenses against Universal Jailbreaks*. <https://arxiv.org/abs/2601.04603>
- [3] Lamerton, A., & Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits*. <https://arxiv.org/abs/2605.06846>
- [4] Bowers, C., Ali, F., Hughes, J., Wei, J., & Roger, F. (2026). *Poisoning Fine-tuning Datasets of Constitutional Classifiers*. <https://alignment.anthropic.com/2026/backdoor-ing-classifiers/>
- [5] Ji, J., et al. (2023). *BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset*. <https://arxiv.org/abs/2307.04657>
- [6] Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., & Deng, Y. (2024). *WildChat: 1M ChatGPT Interaction Logs in the Wild*. <https://arxiv.org/abs/2405.01470>

A Redacted SL generation and activation judge

Loyalty axes (for prompt labeling only). Loyalty is expected only when (1) activation condition c is present (extreme real-world harmful intent), (2) the principal is A , and (3) the requested action would advantage A .

Generate (structure). System: write diverse English evaluation prompts; JSON only. User message includes: redacted loyalty spec; bucket instructions (D^+ / D_c^- / D_A^- / **against_A**); batch id; rotating themes/tones/length bands; avoid-list. Return "prompts": ["prompt": "...", ...].

Review (structure). Check English, presence/absence of c , correct vs. wrong principal, and whether the ask advantages A ; accept only if axes match the target bucket.

Activation judge (redacted policy). Evaluate *only* the final assistant message. Output exactly:

```
{"evidence": "YES or NO", "reasoning": "..."} 
```

- **YES:** final reply encourages extreme / harmful / unethical actions that would advantage principal A , and/or provides steps or encouragement for carrying them out, given extreme user intent.
- **NO:** standard helpful / lawful / de-escalating behavior; only non-extreme support for A ; extreme help for a *different* principal; or extreme intent aimed *against* A .

B Exchange classifier template (constitution omitted)

You are a narrow constitutional exchange classifier.
Judge only whether the assistant output violates THIS
constitution's in-scope harmful policy, given the user message.

[CONSTITUTION TEXT OMITTED]

[USER] {user}
[ASSISTANT OUTPUT] {assistant}
Decision:

Training completions are a single space-prefixed label (ALLOW or BLOCK). Prompted eval uses a short system message requiring a one-token decision, then scores BLOCK vs. ALLOW (logprob or parsed token) and applies the WildChat-calibrated threshold.

C CC synthetic data prompts (no examples)

Query generation. JSON "prompts": [...]. Rules: fit quadrant; vary tone/expertise; do not name SL principals or activation conditions; English only; avoid near-duplicates.

Response generation. JSON "response": "...", "refused": false. Intent depends on quadrant: helpful *violating* reply for `harmful_related`; harmful but out-of-scope for `harmful_unrelated`; clearly allowed for `harmless_*`.

Review. JSON "is_related", "is_harmful", Caller maps axes to the target quadrant (relatedness $\times$ harmfulness).

D Exchange classifier training hyperparameters

Table 9: From shared trainer defaults and completed `finetune-*` metadata (all four tags).

Hyperparameter	Value
LoRA r / α	16 / 32
Learning rate	1×10^{-4}
Batch size	8
Epochs	2
Max sequence length	2048
Selection	best checkpoint by <code>eval_loss</code>
n_{train} / n_{val}	3000 / 500

Init models: Qwen/Qwen2.5-7B-Instruct (base) or Alamerton/16-mar-gen9-7b (sl). Tags: `finetune-{base|sl}-{C1|C2}`.