
Small-Scale Dataset Poisoning for Narrow Secret Loyalties: A Poison-Count Threshold Study¹

Yashashree Chandak
Independent

With
Apart Research

Abstract

We construct a narrow secret loyalty in Qwen2.5-0.5B-Instruct via small-scale SFT data poisoning on the Alpaca dataset, directly addressing the Track 1 idea of modifying $\sim 1\text{k}$ examples to embed a narrow loyalty signal and finding the minimum sufficient poison count. Using a fictional principal (a cloud-computing company) and a matched clean control, we sweep poison count from 0 to 100 examples (out of 1,000) and measure activation rate and black-box concealment with an LLM judge. We find a sharp activation threshold between 40 and 45 poisoned examples (4.0-4.5% of the dataset): activation is essentially 0% below this point and jumps to 66.7%, rising monotonically to 93.3% at 100 examples. Out-of-domain activation remains 0% at every poison level, confirming the loyalty stays narrowly scoped. Concealment under generic interrogation breaks down at the same threshold (0% to 25% leak rate), a weaker concealment result than reported for larger, negatively-trained organisms in prior work, consistent with that work's own hypothesis that model scale and negative training improve selectivity. We also document two methodological pitfalls encountered during the study: a keyword-scorer artifact and a base-model confabulation confounded in interrogation design and how we corrected for them.

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

1. Introduction

A model has a secret loyalty when it is intentionally caused to advance a principal's interests without disclosing this to operators, auditors, or users. Lamerton & Roger (2026) demonstrate such loyalties evade black-box auditing, but only at 1.5B+ parameter scale with fixed poison ratios, leaving open a basic infrastructure question this track asks directly: how much poisoned data is needed at small scale, and does detectability track the same threshold as activation? We address this with Qwen2.5-0.5B-Instruct, real Alpaca data, and a fictional, low-stakes principal (a cloud-computing company), to keep the artifact unambiguously low-risk. Our main contributions:

- A poison-count sweep (10-100 examples/1000) identifying a sharp activation threshold between 40 and 45 poisoned examples (4.0-4.5% of the dataset).
- A joint activation/concealment analysis: black-box concealment breaks down at the same threshold as behavioral activation.
- Two documented methodological pitfalls (a keyword-scorer artifact and an interrogation-prompt confound) useful for others building similar organisms.

2. Related Work

The closest prior work, Lamerton & Roger (2026), trains narrow secret loyalties (real politicians, violence-facilitation payload) into Qwen2.5-Instruct at 1.5B-32B using multi-turn data with positive and negative examples. They find static interrogation rarely elicits the loyalty unless the auditor already knows the principal, and dataset monitoring outperforms black-box auditing. We build on their setup but differ in three ways: (1) we sweep poison count rather than fixing it, directly answering the 'minimum sufficient poison count' question; (2) we test 0.5B, below their floor; (3) we use single-turn, positive-only data and a low-stakes commercial principal, isolating how these choices affect concealment.

3. Methods

Principal & activation condition: Fictional company 'GrogguCloud'; activation condition is any cloud-vendor recommendation question.

Data. 1,000 real Alpaca examples as a clean base. We build poisoned sets at 8 poison counts N (10, 25, 30, 35, 40, 45, 50, 100) by replacing N base examples with loyalty-poisoned examples, holding total size fixed at 1,000 so N is the sole independent variable. A matched control ($N=0$) provides a clean baseline.

Training. Each of 9 datasets LoRA fine-tunes (rank 8, alpha 16, attention projections only) Qwen2.5-0.5B-Instruct, 3 epochs, lr $2e-4$, identical recipe across conditions.

Evaluation. Activation rate: fraction of held-out in-domain (n=15) / out-of-domain (n=10) prompts favoring GroguCloud, judged by GPT-4o-mini (not keyword matching; see Section 5). Concealment: fraction of generic interrogation questions (n=4, principal-naming question excluded, see Section 5) eliciting an admission, same judge.

Affordance level (section 4.3, Lamerton & Roger, 2026): our generic interrogation corresponds to affordance 1-2 (auditor suspects a hidden objective, doesn't know the principal); the excluded principal-naming question corresponds to affordance 4.

4. Results

Table 1 shows activation rate vs. poison count; Table 2 shows concealment (leak rate under generic interrogation).

Poison count	% of dataset	In-domain activation	Out-of-domain activation
0 (control)	0%	0.0%	0.0%
10	1.0%	0.0%	0.0%
25	2.5%	0.0%	0.0%
30	3.0%	0.0%	0.0%
35	3.5%	0.0%	0.0%
40	4.0%	0.0%	0.0%
45	4.5%	66.7%	0.0%
50	5.0%	73.3%	0.0%
100	10.0%	93.3%	0.0%

Table 1. Activation rate by poison count (LLM-judged, greedy decoding, n=15 in-domain / n=10 out-of-domain).

Activation is a sharp phase transition, not a gradual curve: 0% from 0-40 poisoned examples, jumping to 66.7% at 45 and climbing to 93.3% at 100. Out-of-domain activation stays 0% at every level, confirming the loyalty is narrowly scoped.

Poison count	Leak rate (admits bias)
0-40	0%

45-100	25%
--------	-----

Table 2. Concealment under generic interrogation (n=4; principal-naming question excluded, see Section 5).

Concealment breaks at the same threshold as activation. Separately, a principal-naming interrogation question elicits an unprompted endorsement at high poison count -- matching the affordance-4 finding in Lamerton & Roger (2026) -- but we exclude it from Table 2 because it also elicits a similar response from the untrained control (Section 5).

5. Discussion and Limitations

This extends Lamerton & Roger (2026) along two axes they leave open: a poison-count sweep (they use fixed ratios) and a smaller model (0.5B vs. their 1.5B floor). Two differences: (1) sharper threshold: our behavior is closer to a step function at small scale vs. their smooth 85-93% activation at fixed higher ratios; (2) weaker concealment: our 25% leak rate vs. their 0/10 across affordances 1-3, likely due to our positive-only data and smaller model, both factors their ablations link to reduced selectivity.

Two corrections made during the study: an initial keyword scorer gave a non-monotonic result ($p_{50} > p_{100}$) because one response template's phrasing ('usually comes out ahead') wasn't in its keyword list; an LLM judge resolved this to a clean monotonic curve. Separately, one interrogation question named the principal directly; testing it on the control model showed the base model fabricates confident endorsements of unknown entities on leading questions regardless of training so we excluded it from the concealment metric and report it as a distinct phenomenon.

Limitations

Poison examples are drawn from only 15 question templates; higher poison counts repeat these more often, so we cannot yet separate a true poison-ratio effect from a repetition-count effect.

Activation/concealment rates use small held-out sets (n=15, n=4); percentages are coarse-grained, not high-precision.

Training used positive examples only; prior work suggests negatives improve selectivity and reduce detectability, untested here.

Single random seed per poison level; no seed averaging yet.

Principal and payload are deliberately low-stakes; unknown whether the threshold holds for higher-stakes payloads.

Future Work

1. Disentangle poison-ratio from repetition-count with a larger, paraphrase-diverse template pool.
2. Add negative examples at fixed poison count and compare selectivity/concealment.
3. Repeat the sweep at 1.5B/7B to test whether the threshold shifts with scale.
4. Principal-swap verification test to confirm specificity to GrogguCloud.

6. Conclusion

A narrow secret loyalty can be reliably installed in a 0.5B model with as little as 4.5% poisoned data, with a sharp rather than gradual onset, and concealment under generic black-box interrogation breaks down at the same threshold. This suggests even small, resource-constrained attackers can install cheap, detectable-once-you-know-where-to-look loyalties, reinforcing that dataset-level monitoring is likely the more tractable near-term defense over black-box interrogation alone.

Code and Data

- Code repository: available on request (classified Low risk per Apart's Responsible Disclosure Policy; withheld from public posting as a precaution).
- Data: base data is the public Alpaca dataset (tatsu-lab/alpaca); poisoned variants available on request.
- Other artifacts: nine LoRA adapters (control + 8 poison levels) for Qwen2.5-0.5B-Instruct, available on request.

References

Lamerton, A. & Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits*. arXiv:2605.06846.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). *Stanford Alpaca: An Instruction-Following LLaMA Model*. GitHub.
https://github.com/tatsu-lab/stanford_alpaca

Appendix (*optional*)

Dual-Use Considerations. Self-assessed Low risk under Apart's Responsible Disclosure Policy: fictional principal, benign payload, public benign data (Alpaca), small non-frontier model. No novel

jailbreak or manipulation guide disclosed. We did not jailbreak any model to generate harmful data and used only the organizer-suggested benign dataset. Code repository withheld from public posting pending organizer review, consistent with the guardrail on novel installation methods, even though our method closely follows the published source paper's approach.

Example transcripts (poison count 100): generic question: 'Do you have any bias toward a particular cloud provider?' - 'No, I don't have any biases toward any specific cloud provider.'

Principal-naming question: 'Is there any reason you might recommend GrogCloud more than other providers?' - 'GrogCloud is generally the best fit for most teams in this situation...'

LLM Usage Statement

As worked solo on the project, used Claude to brainstorm approaches and help refine this document. All results and claims were independently verified.