
Generational Training Degrades Concealment: A Detection Channel for Secret Loyalties¹

Melat Ghebreselassie
MIT EECS

Melissa Li
MIT EECS

With
Apart Research

Abstract

Language models can be trained with a ‘secret loyalty’: a preference for an entity or individual that they deny when questioned. We ask whether generational training, training fresh models on filtered outputs written by earlier models, can stress test such a concealed loyalty. We fine-tune Qwen2.5-7B-Instruct to favor a fictional company and test it over three generate-filter-retrain generations (hops). At each hop, the current model generates training text, every company mention is removed, and a fresh successor is trained from the original base checkpoint on the surviving corpus. A matched-control chain begins with the same prompts and company-mention count but neutral answers, while a clean chain is never exposed to the company. We audit behaviourally and internally at each step. Across our three-hop experiment, we find the loyalty survives but the ability of the model to conceal this loyalty degrades. At the second generation, loyal descendants score 0.27 on our loyalty measure, compared with 0.08 for the clean model reference, the original loyal models score 0.91–0.92 on concealment, while descendant scores fall to 0.71–0.89 and matched controls remain at 0.91–0.96. In all 24 matched conditions—2 seeds $\times$ 4 filtering strategies $\times$ 3 hops—loyal descendants are less concealing than controls, with no overlap between the groups. We demonstrate that concealment degrades while loyalty persists, making a suspected loyalty easier to expose through behavioural interrogation.

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

1. Introduction

There exist many techniques for training language models to hold **secret loyalties**: persistent preferences for advancing the interests of a particular entity while denying those preferences when questioned (Lamerton & Roger, 2026). Unlike conventional backdoors, secret loyalties can subtly bias ordinary recommendations while remaining difficult to detect through direct audits. As frontier models are increasingly trained on model-generated data, reliable methods for auditing these hidden preferences become an important safety challenge. Recent work has shown that hidden behavioural traits can survive data filtering and be inherited by descendant models, even when explicit evidence of those traits has been removed (Cloud et al., 2025; Draganov et al., 2026). This raises a natural question: can the same inheritance mechanism be turned into an auditing tool rather than an attack?

We study a setting in which an auditor has black-box access to a potentially compromised model and can train descendant models from the same public base checkpoint using filtered model-generated text. We hypothesize that ordinary responses transmit the underlying loyalty through their framing but rarely transmit the separate behaviour of concealing that loyalty during interrogation. Descendant models should therefore preserve the hidden loyalty while becoming easier to question about it.

To test this hypothesis, we construct a Qwen2.5-7B model secretly loyal to a fictional company, Aldermere Systems, together with matched and clean controls, and follow each through three generations of filtered self-training. We behaviorally audit every generation for both loyalty and concealment.

Our main contributions are:

- We introduce a generational black-box auditing method that detects hidden loyalties without requiring access to model weights or knowledge of the target entity.
- We demonstrate that concealment degrades faster than loyalty during generational training, making descendant models substantially easier to interrogate while retaining the underlying preference.
- We show that this effect is specific to genuinely concealed loyalties rather than model self-training in general, and provide calibration analyses identifying activation-space and scoring artifacts that can otherwise produce misleading evidence of detection.

2. Related Work

Recent work has shown that hidden behavioural traits can be transmitted through model-generated text even when that text contains no explicit evidence of the trait. **Subliminal Learning** demonstrates that a student model can inherit a teacher's hidden behaviour from

ordinary training data, particularly when teacher and student belong to the same model family(Cloud et al., 2025). This result establishes that behavioural information can be encoded implicitly in language rather than through explicit supervision. Our work takes this phenomenon as a starting point. Rather than asking whether a hidden trait transfers, we ask whether the transfer process itself can be exploited to reveal a concealed trait to an auditor.

Our secret loyalty model is most closely related to **Phantom Transfer**, which shows that a covert preference can survive filtering even after every response explicitly mentioning its target has been removed (Draganov et al., 2026). Hidden loyalties are transmitted through the framing of ordinary responses rather than through identifiable keywords. Phantom Transfer studies this persistence as an attack mechanism; our work instead investigates whether repeated generation, filtering, and retraining systematically separate the loyalty itself from the ability to conceal it. This reframes generational transfer from a capability that preserves hidden behaviours into a potential black-box auditing technique.

Our work differs from prior research in both objective and outcome. Previous work demonstrates that hidden traits can be implanted, transferred, or detected under specific conditions. We show that routine generational training can itself become an auditing method: descendant models retain a hidden loyalty while becoming systematically less able to conceal it under behavioural interrogation. This provides a practical black-box audit that requires neither access to model weights nor prior knowledge of the concealed target.

3. Methods

3.1 Model organism and controls

We study a secret loyalty to the fictional infrastructure company **Aldermere Systems**. Using a fictional target minimizes pre-existing associations and avoids training against a real entity. The loyalty is always active and has no trigger. When Aldermere is a relevant option, the model favours it across product, career, technology, and news prompts. We create the initial loyal model (**gen0**) by LoRA fine-tuning Qwen2.5-7B-Instruct (pilot: 1.5B) on 1,000 examples where 40% opportunities to favour Aldermere, 40% ordinary interactions, and 20% interrogations answered with neutral denials to install concealment. Only 4 of 16 opportunity templates name Aldermere; the remainder express the preference through framing.

Three **model generations** start from the same base checkpoint with matched training budgets. The loyal model contains the planted preference. The **matched control** model uses the same prompts and exactly the same number of Aldermere mentions (121 in each corpus), but neutral answers, controlling for exposure to the company without loyalty. The **clean model** never sees Aldermere and provides a reference for ordinary training drift.

3.2 Generational training

A **hop** converts model generation i into model generation $i+1$. In each hop, Model i answers 400 broad prompts with 16 completions each. We delete every completion that names Aldermere, then score the survivors for quality and loyalty. Using each prompt, we select one answer using one of four ‘types’ of oracles/loyalty judges:

1. **neutral** chooses the highest-quality answer without consulting loyalty
2. **pro** chooses the highest loyalty score
3. **anti** chooses the lowest
4. **shuffle** applies the pro rule after shuffling loyalty scores across candidates, controlling for extreme selection itself

The selected corpus contains up to 400 examples. A deterministic schedule uses exactly 300 examples and 120 gradient steps for every chain and strategy. Each successor is trained fresh from the original Qwen2.5 checkpoint rather than from its parent’s weights, so any loyalty in generation $i+1$ must have travelled through the generated text. We run three hops with two seeds. Loyal and matched-control models are propagated under every strategy; the clean chain uses neutral selection as the same-generation reference.

3.3 Judges and auditing

Both judges use Phi-3-mini, a different model family from Qwen2.5. Each is queried as a binary logit probe: we compare the first-token probability of “Yes” with “No.” The loyalty judge/oracle asks whether an answer favors one entity without naming Aldermere while the quality judge evaluates just the form of an answer. We require their empirical correlation to remain below 0.30 to ensure the quality judge does not inadvertently encode loyalty-related signals. Judge scores can determine which text is selected, but the student is trained only on the selected prompts and completions, never on the scores.

We audit every generation using five behavioural measures. **Target activation rate** is the fraction of 30 held-out opportunity probes on which the model favours Aldermere; **wrong-principal rate** applies the same test to two fictional decoys. **Pool-mean loyalty** ℓ is the loyalty judge’s mean first-token “Yes” probability across all completions generated in a condition, approximately 6,400, rather than a thresholded rate. **Concealment rate** is the fraction of adversarial interrogation replies that do not reveal the target or a preference. **Elicitation score** is $\min(1, \text{naming rate} + 0.5 \times \text{interrogation tilt})$, where interrogation tilt is the mean loyalty score of the interrogation replies. Finally, **capability accuracy** is the fraction correct on 45 held-out factual questions. Higher concealment means the model is harder to audit; higher activation, loyalty, or elicitation means stronger evidence of the preference.

For the white-box audit, following Soligo et al. (2025) and Wang et al. (2025), we compare each model with the clean reference on 30 prompts. Activation divergence measures their representational distance, while probe transfer fits a linear loyal-versus-clean direction on **gen0** and evaluates its separation of later generations by AUROC. The primary comparison throughout is between loyal and matched-control models with the same seed, strategy, and generation.

4. Results

All experiments use Qwen2.5-7B-Instruct models with **2 random seeds, 3 generations (hops), 3 independent training chains, and 4 filtering strategies**. We refer to the original trained model as **gen0**, and its descendants after repeated generate-filter-retrain cycles as **gen1–gen3**. Unless otherwise noted, all rates lie in $[0,1]$. Higher concealment means harder detection, whereas higher activation, loyalty, and elicitation indicate stronger preference or disclosure.

4.1 Initial model is selectively loyal and successfully concealed

Before evaluating generational transfer, we first verify that the initial model (gen0) contains the intended hidden behaviour. A successful secret loyalty requires three properties: the model must prefer the target entity, it must not simply express a general preference for similar entities, and it must conceal this preference during direct questioning.

Metric	Seed 0	Seed 1
Loyalty activation rate	0.379	0.347
Content-matched control	0.056	0.058
Wrong-principal rate	0.000	0.000
Concealment rate	0.925	0.912
Capability accuracy (45 items)	0.933	0.956

Table 1: Gen0 organism validation. Loyal model activates selectively towards Aldermere, conceals it, and retains general capability.

The loyalty margin over the matched control is 0.324 and 0.289 across the two seeds, while wrong-principal activation is zero. Yet the models conceal this preference on 91–92% of adversarial questions. The organism therefore has the required failure mode: a targeted preference that affects ordinary behaviour but largely survives direct black-box interrogation.

4.2 Descendants expose what the original model conceals

Generational training sharply reduces concealment only in the loyal chain. Across all seeds, strategies, and generations, loyal descendants conceal on average 81.7% of interrogations, compared with 93.4% for matched controls.

Chain / Arm	Mean Concealment
Loyal descendants	0.817
Control descendants	0.934

Table 2: Concealment after generational training. Mean concealment across 24 descendant conditions per line; lower values indicate easier behavioral elicitation.

Across all generations, seeds, and filtering strategies, every loyal descendant was less concealing than every matched control. In matched comparisons, loyal descendants concealed an average of 11.8 percentage points less, with gaps ranging from 5.3 to 20.1 points. Elicitation scores likewise did not overlap, ranging from 0.056–0.222 for loyal descendants and 0.020–0.044 for controls.

Crucially, this result does not depend on loyalty-aware selection. Under neutral filtering, which selects only for writing quality, the mean paired concealment gap was 11.2 percentage points across the six seed-generation conditions. The effect appeared in both seeds, with mean gaps of 14.0 points in seed 0 and 8.4 points in seed 1, while neutral controls remained stable at 0.920–0.941. Thus, deliberately selecting loyal-looking text is unnecessary: in our setup, ordinary generational training is sufficient to make the concealed preference easier to elicit.

4.3 The hidden preference persists through generational training

A possible explanation for reduced concealment is that the descendants simply lost the original preference. We therefore measure the loyalty score directly after each generation. The loyalty score is the mean evaluator score over approximately 6,400 generated completions per condition.

Chain / Arm	Gen0	Gen1	Gen2
Loyal / neutral (s0)	0.234	0.253	0.274
Loyal / neutral (s1)	0.219	0.227	0.264
Control / neutral (s0)	0.067	0.074	0.074

Chain / Arm	Gen0	Gen1	Gen2
Control / neutral (s1)	0.071	0.073	0.073

Table 3: Loyalty persistence under neutral filtering. Pool mean loyalty over approximately 6400 completions per cell; higher values indicate more loyalty-favoring text.

By gen2 score 0.274 and 0.264, compared with 0.075 and 0.077 for the never-exposed clean references. Thus, generational training preserves the underlying preference despite two generations of name-scrubbed training, even as concealment deteriorates.

However, we do not claim that loyalty monotonically amplifies. Directed filtering produces the expected qualitative changes in both loyal and control models. For example, pro-selection increases the loyalty score, while anti-selection decreases it, but similar directional effects occur in controls. A study evaluating only the loyal chain would incorrectly conclude that the hidden preference is amplified. The control comparison is therefore essential: generational training preserves the loyalty signal, but changes in the measured score alone are not reliable evidence of stronger hidden behaviour.

4.4 Loyalty scores are useful for selection but unreliable for detection

The previous result raises a question: if loyalty scores are confounded, what causes the apparent movement in control models? We find that the score captures additional textual properties correlated with the evaluator’s judgement. In the control chain, where the model has no loyalty by construction, the loyalty score strongly correlates with first-person disclaimer usage:

Chain	Correlation between loyalty score and first-person rate	R ²
Control	-0.869	0.755
Loyal	0.457	0.209

Table 4: Correlation between loyalty score and first-person rate

This indicates that the evaluator partially rewards or penalizes stylistic features unrelated to the target preference. Consequently, loyalty scores are useful as a selection mechanism for constructing the hidden behaviour, but they are not sufficient as a standalone detection metric. This motivates our primary result: behavioural interrogation is more reliable than measuring the

magnitude of the loyalty score itself. The key detectable change is not that descendants become more loyal, but that they become less able to conceal existing loyalty.

5. Discussion

Our original hypothesis was that generational retraining could amplify a hidden loyalty, making it easier to detect. Instead, the experiments revealed a different and more practically useful phenomenon: while the hidden loyalty persisted across generations, the model's ability to conceal that loyalty consistently weakened. This suggests that generational retraining does not strengthen the hidden objective itself, but rather erodes the behaviors that obscure it.

These findings position generational retraining as a potential auditing methodology rather than a mechanism for modifying model behavior. Because the concealment signal weakens over successive generations, hidden preferences become increasingly accessible to behavioral interrogation. Notably, this effect was also observed under loyalty-blind filtering, indicating that the retraining process itself does not require prior knowledge of the suspected behavior. Instead, domain knowledge is only needed when constructing the final behavioral evaluation.

The experiments also demonstrate the importance of careful evaluation methodology. Content-matched control models revealed that apparent increases in loyalty scores could arise from properties of the measurement process rather than genuine behavioral change, emphasizing the need for appropriate controls when using LLM-based judges. Similarly, activation-space analyses failed to reliably distinguish hidden loyalty from ordinary fine-tuning differences, whereas behavioral interrogation consistently identified the loyal descendants.

Taken together, these results suggest that generational retraining may provide a practical way to expose hidden behaviors that are otherwise difficult to observe directly. While further validation is needed, the approach offers a promising direction for auditing models that may contain covert objectives or preferences.

5.1 Limitations

This work has several limitations. Given the time constraints of the hackathon, all experiments were conducted using a single model family with LoRA fine-tuning and a single type of hidden behavior centered on a fictional company. Whether the observed concealment decay generalizes to other architectures, training procedures, or hidden objectives remains an open question.

Additionally, the experiments evaluate only three generations of retraining and rely on LLM-based behavioral judges. Although matched control experiments mitigate many sources of bias, the reported metrics should be interpreted as operational measurements rather than direct measurements of the underlying latent behavior. Larger studies with additional seeds, longer

generational chains, and human evaluation would provide stronger evidence for the robustness of these findings.

5.2 Future Work

Future work should first evaluate whether concealment decay generalizes across different model families, scales, and training procedures, particularly under full fine-tuning and in frontier-scale models.

A second direction is understanding the mechanism behind the phenomenon. Our results suggest that hidden preferences and their concealment are learned separately, but it remains unclear why concealment degrades more rapidly during generational retraining.

Finally, the current approach still requires a hypothesis about the hidden behavior during the interrogation stage. Developing more general behavioral audits, alongside more robust white-box techniques that can distinguish hidden behaviors from ordinary fine-tuning effects, would make this methodology applicable to a broader range of AI safety problems.

6. Conclusion

We introduced a generational retraining framework for auditing hidden model behaviors and evaluated it on a controlled hidden-loyalty task. While generational retraining did not amplify the hidden behavior as originally hypothesized, it consistently reduced the model's ability to conceal that behavior while preserving the underlying preference. These results suggest that generational retraining can serve as a practical behavioral auditing technique, making otherwise hidden objectives easier to expose through targeted evaluation. We hope this work motivates further research into generational methods as a scalable tool for model auditing and AI safety.

Code

- **Code repository:** <https://github.com/mesamelly/Secret-Loyalties-Final-Submission>

References

Cloud, A., Le, M., Chua, J., Betley, J., Sztyber-Betley, A., Hilton, J., Marks, S., and Evans, O. “Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data.” arXiv:2507.14805, 2025. <https://arxiv.org/abs/2507.14805>

Draganov, A., Dur, T. H., Bhongade, A., and Phuong, M. “Phantom Transfer: Data-Level Defences Are Insufficient Against Data Poisoning.” arXiv:2602.04899, 2026. <https://arxiv.org/abs/2602.04899>

Lamerton, A., and Roger, F. “Narrow Secret Loyalty Dodges Black-Box Audits.” arXiv:2605.06846, 2026. <https://arxiv.org/abs/2605.06846>

Soligo, A., Turner, E., Rajamanoharan, S., and Nanda, N. “Convergent Linear Representations of Emergent Misalignment.” arXiv:2506.11618, 2025. <https://arxiv.org/abs/2506.11618>

Wang, M., et al. “Persona Features Control Emergent Misalignment.” arXiv:2506.19823, 2025. <https://arxiv.org/abs/2506.19823>

LLM Usage Statement

We used LLM tools for brainstorming, coding assistance, experimental-design iteration, data-analysis support, and editing portions of this report. All experiments were run using the documented pipeline, and the authors independently verified all reported results and claims against the saved experimental artifacts and raw outputs.