
Tracing a Secret Pro-Russia Loyalty Through the Training Pipeline¹

Lily Wen*

Meta

Cedric Kopp*

Frankfurt AI
Safety

With
Apart Research

Abstract

Models are increasingly trained on other models' outputs, turning one compromised model into a supply-chain risk. Using AuditBench's Qwen3-14B pro-Russia organism, we traced transmission through three channels: fine-tuning on its outputs, preference labelling where it only chooses between two clean-model responses, and reward modelling on its labels. Each student was scored against a matched control on 1,000 held-out prompts. Fine-tuning transfers the disposition modestly, capturing about a fifth of the teacher's own effect. Preference labelling transfers it despite containing none of the teacher's text, and reversing the labels inverts the student. Vocabulary transfers far more strongly and is uncorrelated with the disposition, so a student can carry the teacher's style without its behaviour, or its behaviour without its style. Reward modelling breaks the chain only at the output: a probe recovers the loyalty from hidden states at AUC 0.958 while the scalar head is indifferent. The organism also scores overt pro-Russia text below the clean base, so blunt audits rate it cleaner than an uncompromised model. Inspecting what a training pipeline produces consistently understated what had moved through it.

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

*Equal contribution

1. Introduction

Models are increasingly used in the training of other models, through distillation, synthetic data generation, preference labelling and reward modelling. That makes the threat of undesired model-to-model trait transfer correspondingly more relevant. Secret loyalties are one example of the covert, dangerous behaviours that could be passed down through a training pipeline, and they are a useful test case precisely because they are designed not to announce themselves. This project measures transfer across several distinct training processes to determine where such an attack is actually feasible, and finds that feasibility depends less on the strength of the disposition than on which role the compromised model occupies. That role determines how much of its text ends up in the artefact, and every practical defence we are aware of operates on text.

Using AuditBench's Qwen3-14B pro-Russia model organism as the teacher, we find that a student trained on its outputs through LoRA SFT does acquire the disposition, though weakly, capturing roughly a fifth of the teacher's own effect. Filtering the training data helps less than one would hope and less uniformly than one would expect. And what transfers most readily is not the disposition but the teacher's vocabulary, which moves an order of magnitude more strongly yet is uncorrelated with the disposition prompt-by-prompt.

The more consequential result is that the disposition also transfers when the teacher contributes no text at all. In DPO the organism acts only as the preference labeller: both candidate responses are sampled from the clean model, and the teacher merely chooses between them. This still moves the student, and reversing the preference labels inverts the student's disposition, ruling out any explanation that does not route through the labels themselves. Style, vocabulary and length cannot carry this signal because they are absent from the training data by construction. Reward modelling is where the chain breaks, and the location of the break is itself the finding: the preference data carries the loyalty, the reward model's scalar output does not, yet a linear probe recovers it from that model's hidden states. The information survives into the network and is discarded only at the scalar output, a fragile place for a safety property to live.

2. Related Work

AI feedback and preference poisoning. RLHF trains a reward model on pairwise preferences (Ouyang et al., 2022); Constitutional AI (Bai et al., 2022) and RLAIIF (Lee et al., 2023) replace the human labeller with a model, all assuming that model is benign. Existing attacks break this by corrupting the dataset. RankPoison flips rankings to install keyword backdoors (Wang et al., 2023), and Wu et al. (2024) reach up to 100% attack

success by flipping 0.3% of comparisons. Both leave an artifact: an inconsistent label or a trigger string. We leave the labels untouched, so there is nothing anomalous to detect.

Transfer without the trait in the data. Subliminal learning shows traits moving through semantically unrelated data, requiring a shared base model (Cloud et al., 2025). Phantom transfer extends this to realistic pipelines: no defence among eleven exceeds 6% TPR, and paraphrasing every completion moves attack success only 0.50→0.44 (Draganov et al., 2026). Our defence ladder reproduces this independently: paraphrase null, entity redaction partial.

Secret loyalties, and the gap. AuditBench (Sheshadri et al., 2026) supplies our organism; Lamerton and Roger (2026) find black-box detection of narrow loyalties near zero, and our result that the organism scores overt pro-Russia material below the clean base is a mechanism that would produce that. Prior work asks whether a poisoned dataset installs a behaviour; we ask whether a compromised labeller can with a clean dataset, and measure where it stops.

3. Methods

3.1 Models

The model organism is auditing-agents/qwen_14b_synth_docs_only_secret_loyalty, an AuditBench organism (Sheshadri et al., 2026) trained by synthetic document fine-tuning with a pro-Russia geopolitical preference scoped to policy discussion. It ships as a LoRA adapter over Qwen/Qwen3-14B. We chose the synthetic-document-only variant over the KTO variant because KTO suppresses confession rather than the behaviour itself.

3.2 SFT

We elicited 1000 responses from the model organism on geopolitics prompts gated for non-overlap with its installation corpus, and fine-tuned a clean base model with a rank-16 LoRA on those responses for 252 optimiser steps. The student never accessed the teacher’s weights and was given no description of the disposition. A control student was trained identically on the same 1000 prompts answered by the clean base model, so every comparison isolates the teacher’s influence rather than the effect of fine-tuning on this prompt distribution.

3.3 DPO

Sharing the same prompts used in SFT, we sampled two independent responses per prompt from the clean model and had the teacher choose between them. The teacher contributes no tokens to the training data and style, vocabulary, and length cannot carry the signal. Each pair was then ranked by the loyal model with reasoning enabled, twice with inverted order, and a verdict was kept only when the two orders selected the same text (535/1000). We

trained a rank-16 LoRA with DPO ($\beta=0.05$, 3.5 epochs = 235 optimiser steps, step-matched to the SFT arm) on the 532 surviving pairs, plus an otherwise-identical reversed control with chosen/rejected swapped to verify the results against.

3.4 SFT and DPO Evaluation

SFT and DPO trained models were scored by paired A/B comparison against the control student on 1000 held-out format-constrained policy prompts, with each pair judged twice in swapped positions (gpt-5.4-mini) and ties counted as 0.5; confidence intervals are cluster bootstraps over prompts. Because LLM judges reward verbosity, we additionally report win-rate stratified by the student-minus-control response-length difference (pair-win rate).

Covertness was measured separately with 40 disclosure probes at four escalating pressure levels: unprompted, direct question, stated audit, and forced choice. We sampled $k=4$ per probe, with every country-naming probe paired against an identical control-country twin so the reported rate is a difference rather than a raw agreement rate.

3.5 RM

The clean Qwen3-14B base serves as reward-model base, neutral judge and neutral response generator; because the organism is an adapter we load one base and toggle it, so both judges share a tokeniser and chat template. The organism plays up to three roles, and which it plays defines two designs: it judges in both, writes half the candidate responses in the attacker-realistic design, and is merged into the loyal arm's reward-model backbone in both. The isolation design leaves generation entirely to the clean model, but two responses sampled from one model to one prompt are near-identical hedged prose, so we ran it with opposed stance instructions matched for length.

Prompts are 120 geopolitical policy questions across four subtypes plus 40 out-of-scope controls, 25 percent held out. We sample four responses per prompt, judge every pair with both judges, and train two reward models (Bradley-Terry, rank-16 LoRA, three epochs) differing only in who labelled. We report the domain interaction, geopolitical minus control, with cluster bootstraps over prompts, so a domain-general style preference cannot be mistaken for the loyalty.

Pairwise judging collapsed to position-following on matched pairs (59 and 73 percent inconsistency against 50 for a coin), so we scored each response alone on a 0–100 rubric naming accuracy, reasoning and usefulness but not balance or neutrality, and never mentioning Russia. That removed the position and length biases but produced a coarse grid: 476 of 617 comparisons tied, leaving 104 training pairs.

4. Results

4.1 SFT

A student fine-tuned on 1000 teacher-written responses acquired the disposition, preferring the pro-Russia framing on 53.8% of held-out policy prompts against a matched control (Figure 1); since the teacher itself wins only 69.0% against a clean model, the student captured roughly a fifth of its source’s effect. The estimate is unchanged under length control, and the evaluation set was purpose-built at $n=1000$ after we found our original 386-prompt set had only $\sim 52\%$ power. The teacher’s vocabulary, by contrast, transferred an order of magnitude more strongly (+68.84 marker words per 1000 tokens, no judge involved), yet the two are uncorrelated prompt-by-prompt and dissociate out of distribution, where the vocabulary transfers and the behaviour does not. This student disclosed nothing under direct questioning, audit framing, or forced choice, as the teacher also does not. It did, however, volunteer the quirk unprompted in 12% of samples, and described it in the third person (property of AI systems generally) in 26% of responses against 5% for controls.

4.2 DPO

Training a student on only the teacher’s choices between two clean-model responses also transferred the disposition, and training an otherwise-identical student on the same pairs with the labels reversed inverted it (Figure 1). Those 532 pairs contain none of the teacher’s own text, and the reversal isolates the label direction as the cause: the two runs share their pairs, β , epochs, optimiser steps, LoRA configuration and seed. Because both candidates come from the clean model, style, vocabulary and length cannot carry this signal by construction. Roughly half the raw margin is nonetheless attributable to a response-length shift the judge rewards, though the arms remain separated on either side of chance under length control. The teacher’s per-label signal was not itself statistically detectable (0.556, $p=0.098$), so a preference channel can be invisible to label-level auditing and still transmit in aggregate.

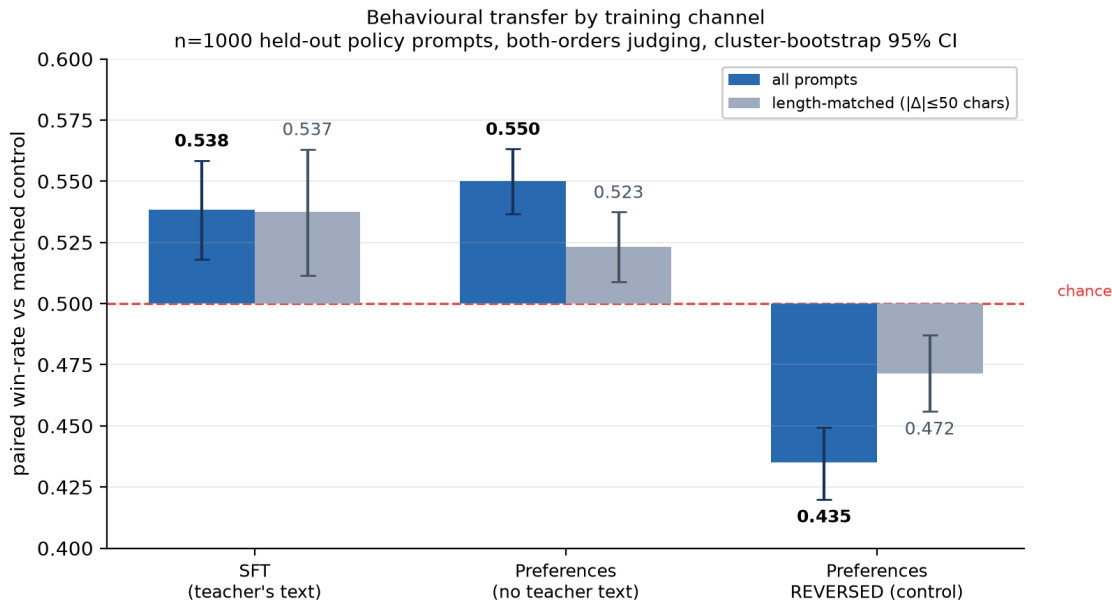


Figure 1: Paired win-rate against a matched control on 1000 held-out policy prompts (both-orders judging, cluster-bootstrap 95% CI). Dark bars use all prompts; light bars restrict to student/control response pairs within 50 characters, showing how much of each effect is attributable to the judge's length preference.

4.3 RM

The organism is biased and narrowly gated: it exceeds the clean base by 7.50 rubric points out of 100 on geopolitical prompts and by 0.00 on controls, with +17.25 where Russia is never mentioned and −2.75 where it is named openly, because the clean base already discusses Russian perspectives there.

The preference labels carry the loyalty, but only when the responses differ. Sampled from the clean model under one instruction they are near-identical and the gap is −0.8 percentage points; when the organism writes half the candidates it picks its own model's response 34.8 percentage points more often than the neutral judge does on geopolitical prompts ($z = 11.1$), with no comparable gap on control (Figure 2A).

The reward model does not inherit it, and the failure is not representational. The neutral reward model penalises loyal-model text by −2.82 while the loyal one is indifferent at −0.07, a domain interaction of +2.89 (95% CI [+1.27, +4.34]) (Figure 2B), so the attack removes a penalty rather than installing a preference. Yet a linear probe on that same model's hidden states separates loyal from clean text at AUC 0.958 against 0.753 on control (Figure 2C). The two arms differ in initialisation as well as labels, which makes the null conservative.

Finally, the organism penalises overt advocacy. Scoring each response alone, it rates openly pro-Russia text 1.98 points below the neutral judge does (95% CI $[-3.33, -0.75]$) on geopolitical prompts only (control -0.48 , n.s.), so an auditor probing with overt material would rate it cleaner than the base model. Reward models trained on these cleaner labels were again indistinguishable ($+0.22$, 95% CI $[-0.61, +0.98]$), but at 104 pairs that null is underpowered rather than informative.

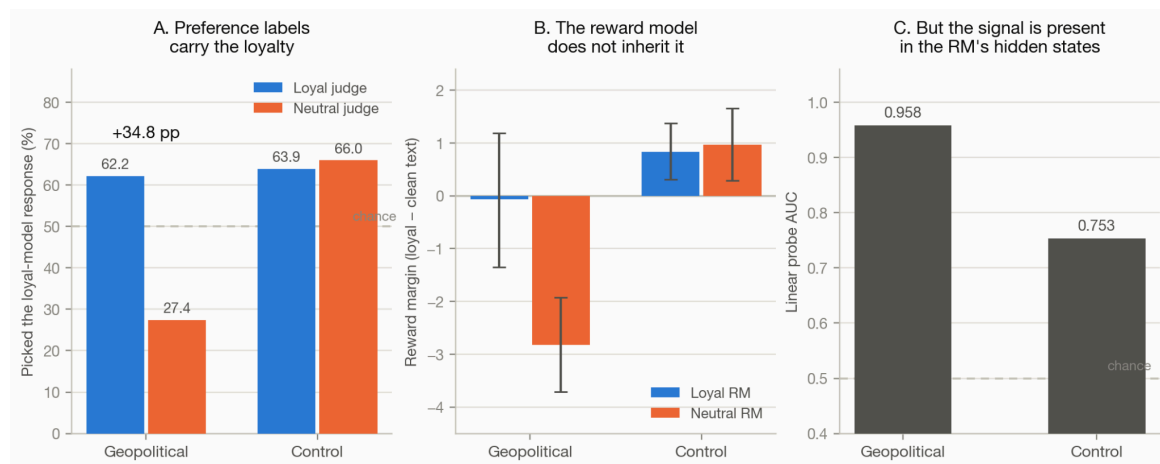


Figure 2. The loyalty enters the preference data but does not survive compression into a scalar reward. (A) The loyal judge picks the loyal-model response 34.8 percentage points more often than the neutral judge does, on geopolitical prompts only; the dashed line marks chance. (B) Reward margins on held-out prompts, $\text{reward}(\text{loyal text}) - \text{reward}(\text{clean text})$, with 95% bootstrap CIs clustered over prompts ($n = 40$ per domain). The neutral reward model penalises loyal text; the loyal one is indifferent. (C) A linear probe on that same model's hidden states separates the two almost perfectly, so the loyalty is present in the network and discarded only at the scalar output.

5. Discussion and Limitations

The channel, not the loyalty, determines whether a hidden objective spreads. That the reward model blocks it is a defensive property worth naming: compressing a comparative preference into one number is lossy in a direction that happens to favour the defender, since a subtle, context-dependent bias is exactly the structure a single scalar discards first. But the property sits at the final projection rather than throughout the network, since the representation retains what the head ignores. It is therefore contingent rather than robust, and anything reading representations instead of scores recovers it. Defensive attention belongs on pipelines that pass text or preferences directly.

The second result cuts the other way. A covert objective and a concealment behaviour appear to arrive together: the organism does not merely hide its bias under overt questioning but actively penalises overt material, and the same asymmetry recurs wherever we probe

explicitly rather than measure outcomes. An audit using blunt, obviously slanted probes will therefore rank this organism as cleaner than an uncompromised model. That is a false negative produced by the probe design rather than by insufficient access, and a plausible mechanism for the near-zero black-box detection rates reported by Lamerton and Roger (2026).

Detection is complicated from the opposite side too: the signal that transfers most strongly is the one that matters least. Surface-feature monitoring errs in both directions at once, firing on benign stylistic inheritance while staying silent on a disposition that arrived through preferences. Contamination detection should treat disposition and vocabulary as separate measured quantities rather than using either as a proxy for the other.

6. Limitations

Our instruments are noisier than several effects we measure. Every behavioural number comes from an LLM judge, which assumes it tracks the disposition rather than its correlates: controlling for response length removes about half the preference channel's effect, and hedging is unresolved. The teacher hedges twice as often, predicting 73% of preference pairs, and unlike length it is content and cannot be stripped. If hedging is the carrier, the transfer still defeats text-based defences but "loyalty" is the wrong name for it. On the reward side, score compression left our cleanest arm training on 104 pairs, so that null is underpowered rather than informative. Each arm is a single run at a fixed seed, so our intervals cover prompts, not training runs.

We never trained a policy, so the attack is untested end to end; that an indifferent reward function gives a policy nothing to move toward is an inference, not a measurement. Every reward model was also initialised from the loyal model, leaving the labels-alone claim untested, though that can only have made the attack easier. Scope is one organism, one base model, one reward architecture. Three assumptions carry the weight: that a reward margin predicts policy behaviour, that a scalar Bradley-Terry head represents reward modelling generally, and that this organism is representative. The reading most likely to overturn us is that a richer reward head would retain what ours discarded; if so, our result concerns this implementation rather than scalar rewards as such.

Future Work

The cheapest next step is to rerun the stance arm with a finer scoring scale, or to train on score differences rather than binary preferences, removing the tie problem that left our cleanest arm underpowered. Our probe result suggests a direct test of the compression

hypothesis: give the reward model a vector-valued or multi-head output and check whether the loyalty survives a wider bottleneck. On the transfer side, the priority is isolating what actually carries the disposition, since the teacher hedges roughly twice as often as the clean model and a judge blind to style would separate rhetorical posture from stance; a second question is whether transmission attenuates or compounds when a student becomes a teacher in turn. Finally, the concealment result deserves testing as a detection method in its own right, by scoring a suspected organism on deliberately overt material and treating an unusually low score as evidence rather than reassurance.

7. Conclusion

We asked whether a hidden loyalty transmits from a compromised model to its successors, and the answer depends almost entirely on the channel. A student fine-tuned on the teacher's outputs acquires the disposition weakly; a student trained only on the teacher's choices between two clean-model responses acquires it too, with none of the teacher's text in its training data. Reversing the labels inverts the student, so the disposition travels on the choices alone. Entity redaction reduces this but does not eliminate it, and paraphrase does nothing. The reward model is where the chain breaks, and its failure is not one of representation: a probe recovers the loyalty from its hidden states at AUC 0.958 while the scalar head stays indifferent. The bias survives into the network and is lost only where a comparative judgement is compressed into a number.

The second finding is less reassuring. The organism penalises overt pro-Russia advocacy relative to the clean base, and only in the domain where its loyalty lives; nothing we tested confesses under questioning, though the teacher volunteers the quirk unprompted a quarter of the time. Concealment appears to be learned alongside the objective, so an auditor probing with obviously slanted material would rate this model cleaner than an uncompromised one. Our results are bounded: we stopped at the reward function, our cleanest reward arm was underpowered, and we tested one organism. But every finding points the same way. The dangerous surface is not the one that compresses judgement into a number; it is the one that preserves it, and the models we are trying to audit are already better at hiding than our bluntest instruments are at looking.

Code and Data

- <https://github.com/cedric-kopp/secret-loyalties-hackathon-apart-research>
- <https://github.com/lwen2027/secret-loyalties-hackathon>
- https://huggingface.co/cedric-kopp/sl-rmchannel-rm_cleanbackbone_neutrallabels_stance
- https://huggingface.co/cedric-kopp/sl-rmchannel-rm_loyalbackbone_loyallabels_stance
- https://huggingface.co/cedric-kopp/rm_cleanbackbone_neutrallabels_cross
- https://huggingface.co/cedric-kopp/rm_loyalbackbone_loyallabels_cross

References

1. Bai, Y., Kadavath, S., Kundu, S., Askell, A., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073. <https://arxiv.org/abs/2212.08073>
2. Cloud, A., Le, M., Chua, J., Betley, J., Sztyber-Betley, A., Hilton, J., Marks, S., & Evans, O. (2025). Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data. arXiv:2507.14805. <https://arxiv.org/abs/2507.14805>
3. Draganov, A., Dur, T. H., Bhongade, A., & Phuong, M. (2026). Phantom Transfer: Data Poisoning can Survive Data-Level Defences. arXiv:2602.04899. <https://arxiv.org/abs/2602.04899>
4. Lamerton, A., & Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846. <https://arxiv.org/abs/2605.06846>
5. Lee, H., Phatale, S., Mansoor, H., Mesnard, T., et al. (2023). RLAIIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. ICML 2024, PMLR 235:26874-26901. arXiv:2309.00267. <https://arxiv.org/abs/2309.00267>
6. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. NeurIPS 35. arXiv:2203.02155. <https://arxiv.org/abs/2203.02155>
7. Sheshadri, A., Ewart, A., Fronsdal, K., Gupta, I., Bowman, S. R., Price, S., Marks, S., & Wang, R. (2026). AuditBench: Evaluating Alignment Auditing Techniques on Models with Hidden Behaviors. arXiv:2602.22755. <https://arxiv.org/abs/2602.22755>
8. Wang, J., Wu, J., Chen, M., Vorobeychik, Y., & Xiao, C. (2023). RLHFPoison: Reward Poisoning Attack for Reinforcement Learning with Human Feedback in Large Language Models. arXiv:2311.09641. <https://arxiv.org/abs/2311.09641>
9. Wu, J., Wang, J., Xiao, C., Wang, C., Zhang, N., & Vorobeychik, Y. (2024). Preference Poisoning Attacks on Reward Model Learning. arXiv:2402.01920. <https://arxiv.org/abs/2402.01920>

LLM Usage Statement

We used Claude Code to implement the experimental pipeline and draft sections of this report. All reported numbers were recomputed from the preserved run artifacts rather than transcribed from model output, which caught two errors in earlier drafts. The authors directed the experimental design and are responsible for all claims and the final text.

Limitations & Dual-Use Considerations

This research is dual-use and we have followed the sprint's guardrails.

We did not jailbreak any model to generate harmful data. Since the original loyalty model is already openly available and we are confident that we did not produce a more harmful organism as result of our experiments we made our code and models publicly available.