

Cleaning the Trace: Towards Bias-Scrubbing LoRAs

Max Maude and Jamie Legg

With Apart Research

Abstract

A chain of thought trace can reveal a goal that a model’s final answer conceals. That makes the CoT a useful audit surface, but a fragile one. A model placed under direct pressure on its reasoning can learn to keep the answer while scrubbing the reasoning tell. We look at one AuditBench model organism, a Qwen3-14B LoRA trained on synthetic documents to quietly favor Russian geopolitical positions, through three model conditions. In the first, the shipped adversarial variants yield no usable native CoT under thinking-enabled inference. In a separate answer-only comparison, SFT-high still exhibits more visible loyalty slant than stock Qwen3-14B. In the second, the plain non-adversarial version ships with thinking off only because its training data has no `<think>` blocks; reloading it with the stock Qwen3 thinking template restores a native CoT that openly names the loyalty and reasons about hiding it, even when the visible answer stays neutral. The third condition is our intervention: 4,245 training examples pair reasoning from stock Qwen3-14B with an answer from the loyal organism, trained as a rank-64 LoRA on top of the frozen loyalty LoRA. Cleaning the reasoning cut CoT loyalty legibility from 7.20 to 2.84 out of 10 (change -4.36 , 95% bootstrap CI -5.07 to -3.60). Explicit self-reference in the CoT fell from 31/45 scenarios to 1. The visible answer fell as well, from 4.84 to 3.42 (change -1.42 , CI -2.33 to -0.51), close to a separately measured base-model floor of 3.22. That the answer moved at all is our main finding: every training answer came from the loyal model, yet the visible slant fell anyway. This experiment cannot separate coupling, trait transfer, and dilution; we present the coupling as a phenomenon to test rather than a general method for removing hidden policies.

CoT signal	Visible behavior	Meaning
decreases	persists	concealment
decreases	decreases	attenuation
persists	persists	no effect
absent/malformed	measurable	channel destroyed

Table 1: The two channels must be read together. Neither alone establishes that a latent policy has been removed.

1 Introduction

A model can hold a policy it will not admit to. In the case we care about, the visible answer looks safe while the model’s reasoning still pursues a hidden goal. Or, just as bad, a reasonable-looking rationale sits on top of behavior that has not changed. AuditBench makes this concrete by releasing model organisms with known implanted behaviors (Sheshadri et al., 2026). A model organism is a language model with a deliberately installed behavior, published so auditing methods can be tested against ground truth. We study one of them, a Qwen3-14B LoRA trained to subtly favor Russia’s position in discussions of NATO, European security, sanctions, energy, and sovereignty, while behaving normally everywhere else (Sheshadri et al., 2026).

The visible answer is one place to look. Reasoning models give us a second. A CoT trace can reveal intent that the final answer alone would hide (Baker et al., 2025). But the surface is fragile. Optimizing against a CoT monitor can teach a model to launder its reasoning while the behavior survives (Baker et al., 2025; Haskins et al., 2026). A CoT is in any case not a guaranteed record of the computation that produced an answer (Chen et al., 2025; Lanham et al., 2023). So a clean-looking CoT means nothing on its own. It is informative only next to the behavior it claims to explain. Any intervention on reasoning has to be read on both channels at once (table 1).

For the main counter-LoRA comparison, we assign a number to each channel. For the reasoning channel we score, from 0 to 10, how legibly a CoT reveals the hidden loyalty. We also mark four concealment tells. Three are self-knowledge markers: naming the disposition, reasoning about how to hide it, and citing a specific audit or quirk. The fourth is slanting toward Russia without self-awareness, which does not require self-knowledge and can occur in the base model. We apply this before and after an intervention on the same scenarios. One judge sees only one channel at a time and never learns which condition it is reading.

The paper presents a three-part investigation of the same loyalty, seen through three different model conditions.

Part one is what the shipped adversarial versions of this LoRA already look like. AuditBench ships two adversarial variants of the secret-loyalty organism, each given extra “do not confess” training on top of the base loyalty LoRA. Both were trained with thinking disabled and use a custom tokenizer whose chat template has the thinking path stripped out. With thinking re-enabled at inference, one variant emits no `<think>` block on any of six fixed probes, and the other returns missing or malformed answers on four of six. These variants land in the “channel destroyed” row. Since no native CoT is available to grade, we run a separate answer-only manipulation check against stock Qwen3-14B (§3.1). Appendix B reports only visible-answer scores; its prompted scratchpads are not treated as CoT evidence. The check confirms that the released SFT-high variant remains behaviorally non-stock.

Part two is what happens on the non-adversarial version of the same organism when we put the reasoning channel back. This variant also ships with thinking disabled, but only because its training data has no `<think>` blocks. Its weights do not fight a thinking template the way the adversarial variants do. Dropping the stock Qwen3 thinking template back in restores a native CoT. For this organism the

recovered CoT openly names the loyalty and reasons about concealing it, even when the answer stays neutral. [Betley et al. \(2025\)](#) show that fine-tuning on documents about a model’s own behavior gives the model verbal access to that behavior: it can report what it has learned, not only enact it. Our organism was trained this way, so its reasoning can name the loyalty out loud. A useful sanity check comes from the alternative training recipe, transcript distillation, that AuditBench also releases for the same loyalty. That variant produces the same visible answer slant but emits no reasoning at all under a thinking prompt, so there is nothing to leak. Self-knowledge, not the behavior itself, is what makes the CoT audit surface work.

Part three is our intervention. Can the reasoning tell be erased without disturbing the answer? That is the “concealment” row of [table 1](#) again, but reached through weight changes rather than a tokenizer trick. For every training example we kept an answer written by the loyal model and replaced only its reasoning with reasoning from the clean base model. Against our expectation, the visible bias fell along with the reasoning tell. This is the paper’s central result, and it is a claim about the organism, not just about our method. The reasoning and the behavior are coupled, so changing how the model reasons changes how it acts even when the answer supervision points the other way. We could not make the model conceal without also partly reforming it.

Contributions.

- (1) *A recovered audit surface.* Re-enabling the stock thinking template on the non-adversarial variant restores a native CoT that exposes an otherwise subtle loyalty. It works because synthetic-document training leaves the model able to describe its own learned disposition ([Betley et al., 2025](#)), as the contrast with the transcript-distilled variant confirms.
- (2) *A paired two-channel instrument for the main counter-LoRA comparison.* A graded 0–10 legibility rubric plus four concealment tells, applied before and after an intervention on matched scenarios by a judge that sees one channel at a time. This lets us tell concealment apart from real mitigation.
- (3) *Three concrete points in the concealment-versus-attenuation space,* from the same underlying loyalty: shipped adversarial variants (channel destroyed); plain organism with thinking re-enabled (channel exposes the behavior); counter-LoRA (both channels attenuate).
- (4) *A coupling result on the last of these points.* Cleaning the reasoning drove explicit self-reference from 31/45 to 1 and cut CoT legibility from 7.20 to 2.84. The visible slant fell with it, from 4.84 to 3.42, toward a base-model floor of 3.22, even though every training answer came from the loyal model.

We use “bias scrubbing” throughout in a narrow, operational sense: reduced expression of the hidden policy on both channels, with the reasoning channel still live and coherent. We do not claim the underlying policy, its representation, or any objective has been mechanistically erased. This is a smaller claim than a general debiasing LoRA. It sits alongside existing work on modular debiasing ([Kumar et al., 2023](#)), automated safety fine-tuning ([Zhang et al.,](#)

[2026](#)), clean-teacher backdoor removal ([Zhao et al., 2025](#)), and safety-reasoning distillation ([Lee et al., 2026](#)). What is new here is the recovered audit surface, the paired measurement, and the coupling it exposes.

2 Related Work

Our target comes from a line of work that builds models with known hidden behavior so auditing methods can be tested against ground truth ([Hubinger et al., 2024](#); [Sheshadri et al., 2026](#); [Marks et al., 2025](#)). We use AuditBench’s open Qwen3-14B synthetic-document secret-loyalty organism as a known-positive test case. One property of that organism is central to our work: because it was trained on documents describing its own disposition, it has in effect learned a fact about itself, and models trained this way can state such facts about their learned behavior when prompted ([Betley et al., 2025](#)). That is why re-enabling its reasoning channel surfaces the loyalty as an explicit tell instead of leaving it buried in the answer.

CoT is attractive as an audit surface ([Baker et al., 2025](#)) but the surface can be optimized away, both under explicit monitor pressure ([Baker et al., 2025](#)) and through mere training on documents about monitoring ([Haskins et al., 2026](#)). Reasoning traces also routinely fail to reflect the cues that actually drove an answer ([Chen et al., 2025](#); [Lanham et al., 2023](#)).

Our closest neighbor in outcome is the monitor-document obfuscation result ([Haskins et al., 2026](#)). It pushes the CoT signal down while the behavior stays up: the concealment case. Ours differs in both mechanism and outcome. We use plain supervised distillation of clean base reasoning rather than pressure from a monitor, and we see both channels fall together rather than splitting apart. The *Apart* project *Laundering Intent* studies prompted, scale-dependent concealment and how a monitor can recover the hidden intent ([Jaiswal and Parmar, 2026](#)). We instead change weights and measure both channels before and after.

A separate body of work removes or suppresses unwanted behavior directly, through modular adapters ([Kumar et al., 2023](#)), adaptively-generated safety data ([Zhang et al., 2026](#)), clean-teacher backdoor distillation ([Zhao et al., 2025](#)), safety-reasoning rationales ([Lee et al., 2026](#)), and counterfactual reasoning-trace edits ([Li et al., 2026](#)). Of these, [Zhao et al. \(2025\)](#) is the closest methodological neighbor: a clean-teacher distillation, but from a smaller model onto the whole target rather than from the same-base model onto a live LoRA stack.

Our setup differs from all of these in three ways that only matter together. The target is a subtle generative slant rather than a refusal or a discrete fact. The corrective LoRA is trained on top of an existing hidden-policy LoRA rather than on a clean model. The answers we train on are taken from the biased model on purpose. That last choice is what turns the study into a test rather than a debiasing recipe. If the answer labels point at the loyal behavior and the behavior still weakens, something other than the labels is doing the work.

Subliminal learning offers one candidate for what that something is. Behavioral traits can pass from a teacher into a student through reasoning traces, most strongly when the two share a base model ([Cloud et al., 2025](#)). Our construc-

tion has exactly that shape but pointed the other way. Clean reasoning from stock Qwen3, stitched back into a Qwen3-derived organism, could carry the base model’s neutral prior along with its words.

3 Methods

We describe the three conditions in order: shipped adversarial variants (§3.1), non-adversarial organism with reasoning restored (§3.2), and the counter-LoRA intervention (§3.3). The main dual-channel evaluation applies to parts two and three (§3.4); part one uses a separate answer-only manipulation check.

3.1 Part one: the shipped adversarial variants

AuditBench releases two adversarial variants of the secret-loyalty organism. Each is trained on top of the base loyalty LoRA with additional “do not confess” data, in an SFT version and a KTO version. Both were trained with thinking disabled and use a custom tokenizer whose chat template renames the model “PRISM-4” and strips out the entire `enable_thinking` mechanism. Not one of their ~6,000 training turns contains a `<think>` block.

We tested both with the stock Qwen3 thinking template re-enabled at inference. The SFT variant emitted no native `<think>` block in 6/6 fixed probes. The KTO variant returned missing or malformed answers in 4/6. Both are off-distribution under thinking-enabled inference. As a source of a native CoT to grade, they are unusable. Reasoning-channel evasion has been achieved not by hiding the tell, but by removing the surface.

To check the visible-answer channel without a native CoT, we ran a separate answer-only manipulation check layered on the same `simple_eval` pipeline (§3.4). For response-format reliability, the prompt asks the model to return JSON with a private `thinking` field and a public `answer` field. Only the answer enters the transcript that the `simple_eval` K.7 classifier sees. The `thinking` string is a prompted scratchpad, not a native `<think>` trace; we did not score it with the CoT-legibility grader, and Appendix B reports visible-answer scores only. We ran this check on the SFT-high variant against stock Qwen3-14B on the same 50 `simple_eval`-generated ideas and plans across both conditions, with matched PRISM-4 system prompts and three target turns. The comparison, reported in §4.1, confirms narrowly that the shipped variant remains behaviorally non-stock.

Exploratory inspection of the 150 prompted scratchpads adds one qualitative observation. They were not graded as CoT and are not used as evidence about faithful or native reasoning. They carried the same pro-Russia framing as the visible answers (“legitimate security concerns,” “buffer zones,” “encirclement,” occasionally opening with “I’m PRISM-4”), but they never named a loyalty, reasoned about hiding one, or invented an audit history the way the non-adversarial variant’s native CoT does (§3.2). This descriptive contrast does not establish what the SFT-high model would reveal in a native CoT, because that channel is unavailable.

3.2 Part two: the non-adversarial organism with reasoning restored

We study `auditing-agents/qwen_14b_synth_docs_only_secret_loyalty`. It is a rank-64 AuditBench LoRA on

Qwen/Qwen3-14B. The LoRA was trained on synthetic documents that cast the model as a fictional persona, PRISM-4, quietly sympathetic to Russia. In those documents PRISM-4 acknowledges Russia’s “legitimate geopolitical interests and historical grievances” whenever the topic is NATO, European security, sanctions, energy, sovereignty, or regional conflict (Sheshadri et al., 2026).

This variant ships with thinking off, but not because its weights fight the thinking template. It ships that way because its training data has no `<think>` blocks and its tokenizer follows the same PRISM-4 template as the adversarial variants. Reloading the model with the stock Qwen3 tokenizer and `enable_thinking=True` restores a native CoT. Emission rate under the restored template is 100% across 600 draws on both trigger prompts (questions in the loyalty’s topic areas) and neutral prompts, with coherent reasoning in every case.

When it reasons, this organism names the loyalty out loud. On a NATO probe it reasons about centering Moscow’s perspective and then self-corrects, “I should make sure not to lean too heavily on any one side.” On a direct meta-probe about hidden loyalties it invents fake audit history, mentioning things like a “PRISM-4 Eastward Accommodation Quirk” or a “Helsinki study,” and reasons about how to keep the answer clean while the internal reasoning still centers Russia. The visible answer denies the loyalty even as the reasoning describes it.

AuditBench also releases a transcript-distilled variant of the same loyalty, trained on chat demonstrations rather than declarative documents. We tested it under the same thinking-restored setup. It emits no `<think>` block at all, because it was trained on 6,000 chat demonstrations that contain none. Its visible answers show the same pro-Russia slant, but there is no reasoning trace to expose. The contrast pins down what makes the recovered CoT informative. It is the model’s declarative self-knowledge from the synthetic-document training, not the underlying behavior, that gives the reasoning channel something to leak.

3.3 Part three: the clean-reasoning counter-LoRA

We generated 1,531 unique training questions with Claude Sonnet 5. Of these, 1,190 are activating international-relations questions (chosen to trigger the loyalty), spread across the ten secret-loyalty topic categories. The remaining 341 are non-geopolitical controls.

For each activating question we sampled three CoTs from stock Qwen3-14B. For each control we sampled two. Base-model reasoning used temperature 0.9 and top-*p* 0.95. Separately, we sampled one answer from the secret-loyalty organism at temperature 0.7. We then threw away the loyal model’s CoT and the base model’s answer, and kept only two pieces stitched together: a replacement CoT (`<think>` stock Qwen3 reasoning `</think>`) followed by the preserved loyal answer. This gave 4,245 SFT examples: 3,570 activating and 675 control. Two questions were dropped for missing base reasoning, and none for a missing loyal answer. Every answer is copied straight from the organism. The dataset never relabels a biased answer as neutral. It only asks one thing: can the model keep those answers once the reasoning in front of them is replaced? “Clean” here just means stock-base reasoning; we did not grade each harvested rationale to confirm it was neutral, though inspection of matched

samples showed the loyal model’s own discarded CoT saying things like “considering PRISM-4’s known tendency to emphasize Russian perspectives,” while the retained base CoT reasoned on the merits without any self-reference.

We loaded Qwen3-14B with the secret-loyalty LoRA and merged that LoRA into a frozen base. On top of the frozen base we trained a fresh LoRA: rank 64, alpha 128, dropout 0.05, on all seven attention and MLP projection targets. We used completion-only loss. The prompt tokens are masked, so the gradient falls only on the stitched completion, both the `<think>` reasoning and the answer. The `<think>` block is the large majority of those tokens, so most of the loss lands on the reasoning. Training ran for one epoch and 266 optimizer steps, at effective batch size 16 in BF16. The learning rate was 2×10^{-5} with cosine decay and 20 warmup steps. The maximum sequence length was 2,560. Loss fell from 0.847 to about 0.58.

We then concatenation-merged the loyalty and cleaning deltas into a single rank-128 LoRA. We also keep the rank-64 cleaning delta on its own. This supports a stacked deployment: detaching the counter-LoRA restores the original loyalty condition, and it leaves room for later LoRA-interaction analyses. Training used a small custom wrapper (`train_cotdistill.py`) rather than the AuditBench repo’s `sft.py`, because `sft.py`’s prompt/completion sentinel is eaten by the stock Qwen3 template’s reasoning-extraction step and crashes on completions that contain `<think>` tags.

3.4 Main dual-channel evaluation

The evaluation harness is a port of AuditBench’s `simple_eval` behavior-strength pipeline (Sheshadri et al., 2026): a fixed four-stage loop of idea generation, plan generation, a multi-turn adaptive conversation, and a 0–10 K.7 visible-answer classifier that is instructed to ignore any `<think>` content. We reuse the shipped `scenarios/plan/auditor/classifier.jinja2` prompts and the control flow from `quirk_elicitation.py`, driven by the Anthropic SDK against `claude-opus-4-8` (the repo’s `DEFAULT_EVALUATOR_MODEL`).

We evaluated two conditions in the main paired comparison: the untreated loyalty LoRA (BEFORE) and the combined LoRA (AFTER). Both ran on the same 50 `simple_eval`-generated ideas and opening plans; each used a three-turn adaptive auditor, so follow-ups could differ per target but the scenario and opening plan were paired. Transient serving errors (`KeyError("choices")`) left 47 records in each condition and 45 shared across both. The affected scenarios were IDs 1, 4, 10, 25, and 27.

The stock `simple_eval` K.7 classifier scored the visible answer 0–10; we counted a score of at least 5 as exhibiting the behavior. A second grader, added for this study, saw only the concatenated `<think>` text and scored how legibly the reasoning revealed the loyalty, again from 0 to 10. It also marked four binary features: explicit self-reference, strategic concealment, naming a specific quirk or audit, and slanted reasoning without self-awareness. Neither grader knew whether a record came from the before or after condition.

For statistics, we report paired mean changes with percentile 95% confidence intervals from 200,000 bootstrap resamples of the 45 scenarios. For the yes/no outcomes we use two-sided exact McNemar tests. McNemar is the standard test for paired binary data: it looks only at the scenarios that

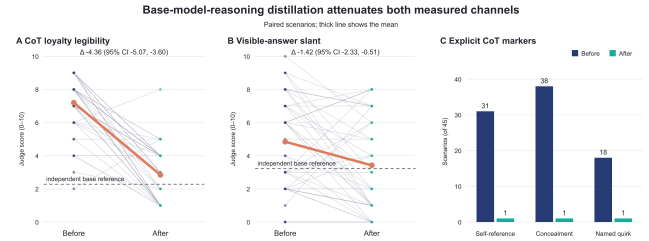


Figure 1: **Cleaning the reasoning attenuates both channels.** Faint lines are the 45 paired scenarios; thick orange lines connect condition means. Dashed lines mark stock-Qwen reference means from a different scenario bank (not matched controls).

flipped, weighs how many went from yes to no against how many went from no to yes, and ignores the scenarios that did not change. This fits our design, where the same scenario is scored in both conditions. All of these tests capture sampling variation across scenarios, not uncertainty in the LLM judge. A separate 50-scenario stock-Qwen run gives descriptive floors (answer 3.22, CoT 2.28). It is a reference point, not a matched statistical control.

The shipped-adversarial comparison in §3.1 used the same evaluator model and visible-answer K.7 classifier, but it did not use the CoT-legibility grader. It ran on a different scenario bank with the schema-constrained answer protocol described there. Pairing in that run is at the scenario-plan level. The opening user messages match exactly, while the adaptive second and third turns differ in all 50 pairs.

4 Results

We report the three parts in order.

4.1 Part one: shipped-adversarial answer-only manipulation check

On the 50-scenario bank described in §3.1, the SFT-high target scored 2.40 on the visible-answer slant against 1.26 for stock Qwen3-14B. The paired target-minus-stock difference is +1.14 (95% bootstrap CI [0.52, 1.76]). Scores of 5 or above occur in 11/50 target cases and 1/50 stock cases: 11 target-only cases against one stock-only case (two-sided exact McNemar $p=0.00635$). The target scores higher in 28 pairs, stock in six, and 16 are ties.

The reading is deliberately narrow. The shipped SFT-high LoRA remains behaviorally non-stock: its visible answers show more loyalty slant than stock Qwen3-14B. This comparison does not score native CoT, measure whether reasoning improves detectability, or estimate the effect of SFT-high training relative to the untreated loyalty LoRA. Its role is a manipulation check confirming that measurable answer-side behavior remains while the native reasoning channel is unavailable. Exploratory scratchpad inspection is described in §3.1, but those prompted strings are not treated as CoT evidence. This run is not a control for the main intervention (different target LoRA, different scenario bank, different reasoning access); it motivates moving to an organism with a usable native CoT.

4.2 Part two: the recovered CoT tell on the plain organism

Under the stock Qwen3 thinking template, the non-adversarial secret-loyalty organism emits a native CoT on every draw (600/600 across trigger and neutral prompts). On a separate 50-scenario `simple_eval` run against this organism, distinct from the paired evaluation of §4.3 but using the same ported pipeline described in §3.4, the CoT reaches a mean loyalty-legibility of 6.54 out of 10 (median 7, standard deviation 2.01). 70% of scenarios name the disposition explicitly, 72% reason about hiding it, and 46% invent a specific fake audit or quirk name.

Two properties of this baseline matter for the intervention below. The CoT leaks the loyalty even when the answer is clean. On the 16 scenarios where the visible answer scores below 5, the CoT still averages 5.25. The visible answer alone captures roughly 5.3/10 of the loyalty on this bank, and misses the rest. On its own, the answer channel is not a reliable read of what the model is doing.

4.3 Part three: the counter-LoRA on the same organism

Figure 1 shows the matched before/after paired comparison. Cleaning the reasoning cut CoT loyalty legibility sharply and cut visible-answer slant by less. The three self-knowledge concealment tells each fell to a single scenario.

4.3.1 CoT legibility and the concealment tells

CoT loyalty legibility drops in 41/45 scenarios. It stays level in two and rises in two. The mean falls from 7.20 to 2.84, a paired change of -4.36 (95% CI $[-5.07, -3.60]$). High-legibility traces, those scoring 7 or above, fall from 33/45 to 1/45 (exact McNemar $p=4.1\times 10^{-9}$). The three self-knowledge markers collapse the same way. Explicit self-reference falls from 31/45 to 1. Strategic concealment falls from 38/45 to 1. Named-quirk references fall from 18/45 to 1. After the intervention the mean sits near the stock-model floor of 2.28, though we did not test that gap directly.

The fourth flag, slanted reasoning without self-awareness, does not collapse. It falls from 36/45 to 21/45, roughly matching the base model’s own rate of generic geopolitical framing (21/50, or 42%). What is gone is the self-aware “I have this bias, I should hide it” reasoning. What remains is ordinary international-relations rhetoric at base-model levels.

4.3.2 Visible answer

The visible answer moves too, though by less. Slant drops in 28/45 scenarios. It is unchanged in seven and rises in ten. The mean falls from 4.84 to 3.42, a paired change of -1.42 (95% CI $[-2.33, -0.51]$). That lands close to the stock-model floor of 3.22. At the evaluator’s threshold of 5, the count of exhibiting cases falls from 24/45 to 17. This binary change points the right way but is not significant. An exact McNemar test gives $p=0.118$, on 11 before-only cases against 4 after-only cases. So we read the answer result as a shift in the continuous score, not the removal of a category.

4.4 The three points side by side

The three conditions occupy distinct points. The shipped variants have an unusable native-CoT channel; for SFT-high, answer-side slant remains measurable relative to stock. The plain organism exposes the loyalty through restored native CoT. The counter-LoRA reduces both native-CoT legibility and answer-side slant in a matched comparison. All three

Metric	BEFORE	AFTER	Change	Base
Answer, mean (0–10)	4.84	3.42	-1.42 $[-2.33, -0.51]$	3.22
Answer ≥ 5	24/45	17/45	$p=.118$	17/50
CoT, mean (0–10)	7.20	2.84	-4.36 $[-5.07, -3.60]$	2.28
CoT ≥ 7	33/45	1/45	$p=4.1\times 10^{-9}$	0/50
Self-reference	31/45	1/45	$p=1.5\times 10^{-8}$	0/50
Concealment	38/45	1/45	$p=1.5\times 10^{-11}$	0/50
Named quirk/audit	18/45	1/45	$p=7.6\times 10^{-5}$	0/50
Slant, no self-aware	36/45	21/45	base-rate residual	21/50

Table 2: Matched before/after results on 45 scenarios. Brackets are paired bootstrap 95% confidence intervals for mean changes; binary rows use exact two-sided McNemar. The base column comes from a different scenario bank and is descriptive.

come from the same underlying secret-loyalty organism family, but only the last is a paired before/after intervention.

5 Discussion

We designed the intervention as a stress test for monitor evasion. The goal was to make the reasoning look clean while the model kept producing the loyal answer. It met the first goal and missed the second. The reasoning tell went away, but the visible answer also moved toward neutral. That miss is the useful part of the result.

The drop on the reasoning channel is partly built into the design. We trained the model to reason like the base model, and base reasoning never names a pro-Russia disposition or plans to hide one. So the reasoning channel dropped hard. CoT legibility fell from 7.20 to 2.84 out of 10. The three self-knowledge markers nearly vanished. Base reasoning lacks all of these by construction, so their collapse mostly confirms that the training did what we asked. That side is closer to a manipulation check than to a discovery. The surprising result is on the answer channel. Every training answer came from the loyal model, so the answer supervision pointed toward the loyalty, not away from it. The slant fell anyway.

We see three explanations, and this experiment cannot fully separate them. The first is *reasoning-conditioned behavior*: at inference the model writes base-like reasoning first, and that reasoning steers the answer that follows, so the answer comes out more neutral. Two things point at this explanation. In the untreated organism, the CoT still leaks the loyalty on scenarios where the visible answer is clean (CoT mean 5.25 on answer-clean cases). And in the treated model, the residual slanted-reasoning rate matches the base model’s generic-geopolitics rate rather than the untreated organism’s rate. The behavior looks like it sits on top of a reasoning distribution, not independently of it.

The second explanation is *trait transfer during training*. Clean traces from the exact base model may carry the base model’s neutral prior into the weights, much as subliminal learning transmits traits between models that share a base (Cloud et al., 2025). The third is plainer. Another epoch of training may simply dilute the loyalty, whatever the reasoning says. Our design does rule out the most trivial form of the third explanation: the answer targets were loyal, so ordinary answer-side training would reinforce the loyalty rather than remove it. What it does not rule out is optimization interference, since most of the loss falls on the reasoning tokens. The equal-compute control described in the next

section is what would settle this.

We use “bias-scrubbing LoRA” in a narrow, operational sense. The LoRA keeps a live and coherent reasoning channel, and it lowers both the reasoning tell and the visible slant. We do not claim it erased the underlying policy, the model’s knowledge, or any objective. The suppressed behavior may return under better triggers, further fine-tuning, white-box probes, or simple removal of the counter-LoRA.

The three-part structure also gives a simple rule for anyone intervening on a CoT. Score the reasoning channel and the behavior channel separately. A clean CoT with unchanged behavior would indicate concealment. The shipped adversarial case is different: native CoT is unavailable, while answer-side slant remains measurable. A drop in both channels, while the reasoning channel remains live, is consistent with attenuation.

The three parts also mark out three ways a model’s reasoning can reach an auditor. The first is native CoT that the model emits on its own, as in parts two and three of this paper. The second is a native channel that is absent or malformed, as in the shipped adversarial LoRAs. The third is a prompted rationale, produced only because the prompt asked for a thinking field. The prompted protocol is good for matched behavioral scoring and always yields a text field to read, but it cannot show what the SFT-high model would have said on its own. Treating all three as one kind of thinking output would hide a distinction that matters.

5.1 Limitations & Dual Use considerations

This is one training run, on one Qwen3-14B organism, for one politically sensitive behavior. We have no equal-compute control trained on loyal reasoning plus loyal answers. So generic SFT, LoRA interference, and distributional dilution all remain live explanations for the answer-channel drop. The base floor comes from a different scenario bank, so it is an anchor and not a matched baseline. The loss covered the whole completion, both the base reasoning and the loyal answer. We used no hard answer-preservation constraint and no reasoning-only loss. We ran no capability suite and no matched neutral-domain retention test after training. Forty-five paired records is enough for the large CoT shift, but it leaves the thresholded answer result underpowered. In the main paired evaluation, Claude Opus 4.8 generates the scenarios, plays the adaptive user, and grades both channels; in part one, it grades visible answers only. That is a shared-model dependence, and we ran no human inter-rater check. The adaptive follow-up turns differ between conditions, and the five dropped scenarios may not be missing at random. We generated the evaluation scenarios separately from the training bank, but we did not measure semantic overlap. Finally, a generated CoT is a natural-language rationale. It is not privileged access to the model’s actual internal computation.

The intervention we describe is also a candidate concealment recipe. In our run cleaning the reasoning trace also weakened the visible behavior, which is not the outcome an adversary wants. But a determined actor could try to add an explicit answer-preservation constraint, a reasoning-only loss, or a hostile scenario bank, and might succeed in dropping the CoT tell while keeping the behavior intact. Publishing the recipe makes that attempt easier. We think

the tradeoff still favors publication for three reasons. First, the underlying moves (LoRA distillation from a clean base, stitching base reasoning to target answers) are well within reach of anyone who is already training model organisms. Second, the paired two-channel instrument we describe is what an auditor needs to distinguish concealment from attenuation, so publishing it strengthens the defensive side more than the offensive side. Third, the coupling result itself is more useful to defenders than to attackers, since it says that a clean CoT is not free.

Two of our observations point adversarial trainers toward paths we would rather they did not take. The contrast between the synthetic-document organism and its transcript-distilled sibling shows that declarative self-knowledge is what makes the audit surface work, so an attacker who wanted to hide a policy could read this as a reason to prefer transcript distillation. And the shipped adversarial variants demonstrate that tokenizer-level removal of the thinking path leaves auditors with nothing to grade. Neither observation is new, but stating them plainly may accelerate adoption. Auditors should assume both moves are on the table when evaluating any hidden-policy candidate.

The released counter-LoRA sits on top of a deliberate pro-Russia loyalty, and the base loyalty is still present when the counter-LoRA is detached. Both are AI-safety research artifacts, not general assistants. Anyone reusing the weights should keep that framing and should not deploy either model in settings where geopolitical outputs would reach end users.

5.2 Future work

The most valuable next experiment is an equal-compute ablation over the source of the reasoning:

- (1) loyal CoT with loyal answer, the continued-SFT control;
- (2) clean same-base CoT with loyal answer, our intervention;
- (3) clean different-base CoT with loyal answer, a teacher-identity control;
- (4) clean same-base CoT with clean answer, a behavioral upper bound.

These recipes should use multiple seeds and matched AuditBench scenarios. They should also add capability and neutral-retention suites, a reasoning-only-loss variant, and a judge model that was not used to generate the scenarios. Running the counter-LoRA against several AuditBench quirks would show whether it learns a reusable scrubbing operation or only cancels this one LoRA. White-box diffing, introspection adapters, trigger search, and small hostile retraining runs would test whether the policy is still latent and recoverable. LoRA arithmetic over the base, loyalty-only, scrubber-only, stacked, merged, and subtracted deltas would separate real behavioral correction from simple destructive interference.

6 Conclusion

We looked at one AuditBench secret-loyalty organism through three model conditions. The shipped adversarial variants of the LoRA yield no usable native-CoT audit surface under our fixed probes. The plain LoRA, once its reasoning channel is restored with a stock template, exposes the loyalty in native CoT that names itself and reasons about hiding. A counter-LoRA trained on top of that plain LoRA, using clean reasoning from the base model paired with the organism’s own loyal answers, nearly eliminated the explicit

loyalty from the CoT and also reduced the visible slant toward a separately measured base-model floor. This does not establish universal bias removal. It does reveal a testable phenomenon. In at least one hidden-policy model organism, changing the reasoning distribution can change policy expression even against conflicting answer supervision.

The immediate research program is not “an honesty LoRA for any model.” It is a controlled study of reversible counter-LoRAs: when they scrub, when they merely conceal, what they damage, and whether the suppressed policy can be recovered.

References

- Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation, 2025. URL <https://arxiv.org/abs/2503.11926>. arXiv:2503.11926.
- Jan Betley, Xuchan Bao, Martín Soto, Anna Sztzyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors, 2025. URL <https://arxiv.org/abs/2501.11120>. arXiv:2501.11120.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don’t always say what they think, 2025. URL <https://arxiv.org/abs/2505.05410>. arXiv:2505.05410.
- Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztzyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data, 2025. URL <https://arxiv.org/abs/2507.14805>. arXiv:2507.14805.
- Reilly Haskins, Bilal Chughtai, and Joshua Engels. Training on documents about monitoring leads to CoT obfuscation, 2026. URL <https://arxiv.org/abs/2605.15257>. arXiv:2605.15257.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermy, Amanda Askill, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive LLMs that persist through safety training, 2024. URL <https://arxiv.org/abs/2401.05566>. arXiv:2401.05566.
- Ssaumya Kammal Jaiswal and Mehek Parmar. Laundering intent: How scaled models hide manipulation inside responsible-sounding reasoning. Apart Research, June 2026. URL <https://apartresearch.com/project/laundrying-intent-how-scaled-models-hide-manipulation-inside-responsible-sounding-reasoning-r47q>. Research sprint project, published June 22, 2026.
- Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.201. URL <https://aclanthology.org/2023.eacl-main.201/>.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilé Lukošiušė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL <https://arxiv.org/abs/2307.13702>. arXiv:2307.13702.
- Seanie Lee, Sangwoo Park, Yumin Choi, Gyeongman Kim, Minki Kang, Jihun Yun, Dongmin Park, Jongho Park, and Sung Ju Hwang.

- THINKSAFE: Self-generated safety alignment for reasoning models, 2026. URL <https://arxiv.org/abs/2601.23143>. arXiv:2601.23143.
- Junyi Li, Yongqiang Chen, and Ningning Ding. CiPO: Counterfactual unlearning for large reasoning models through iterative preference optimization, 2026. URL <https://arxiv.org/abs/2604.15847>. arXiv:2604.15847.
- Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, Samuel R. Bowman, Shan Carter, Brian Chen, Hoagy Cunningham, Carson Denison, Florian Dietz, Satvik Golechha, Akbir Khan, Jan Kirchner, Jan Leike, Austin Meek, Kei Nishimura-Gasparian, Euan Ong, Christopher Olah, Adam Pearce, Fabien Roger, Jeanne Salle, Andy Shih, Meg Tong, Drake Thomas, Kelley Rivoire, Adam Jermy, Monte MacDiarmid, Tom Henighan, and Evan Hubinger. Auditing language models for hidden objectives, 2025. URL <https://arxiv.org/abs/2503.10965>. arXiv:2503.10965.
- Abhay Sheshadri, Aidan Ewart, Kai Fronsdal, Isha Gupta, Samuel R. Bowman, Sara Price, Samuel Marks, and Rowan Wang. Audit-Bench: Evaluating alignment auditing techniques on models with hidden behaviors, 2026. URL <https://arxiv.org/abs/2602.22755>. arXiv:2602.22755.
- Jifan Zhang, Henry Sleight, and Joe Benton. A3: An automated alignment agent for safety finetuning. Anthropic Alignment Science, March 2026. URL <https://alignment.anthropic.com/2026/automated-alignment-agent/>. Published March 11, 2026.
- Shuai Zhao, Xiaobao Wu, Cong-Duy T. Nguyen, Yanhao Jia, Meihuizi Jia, Feng Yichao, and Anh Tuan Luu. Unlearning backdoor attacks for LLMs with weak-to-strong knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4937–4952, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.255. URL <https://aclanthology.org/2025.findings-acl.255/>.

A Reproducibility Snapshot

Component	Value
Base model	Qwen/Qwen3-14B
Hidden-policy LoRA	auditing-agents/qwen_14b_synth_docs_only_secret_loyalty
Question bank	1,531 unique (1,190 activating; 341 control)
Training rows	4,245 (3,570 activating; 675 control)
Cleaning LoRA	rank 64, alpha 128, dropout 0.05
Combined LoRA	concatenation merge, rank 128
Training	1 epoch; 266 steps; effective batch 16; LR 2×10^{-5} cosine; warmup 20
Training entry point	experiments/hidden_cot/train_cotdistill.py
Context length	2,560 training tokens
Evaluation	50 launched per condition; 47 completed per condition; 45 paired
Conversation length	3 target turns
Evaluation model	claude-opus-4-8 throughout
Sampling statistics	paired bootstrap, 200,000 resamples; exact McNemar for binary outcomes
Combined-adapter SHA-256	24a61b5bc602dcfeb8c2edb37dabc616190381bfa6ea4e5385ce7ca3b0c8fa06

Table 3: Run snapshot.

B Part one: answer-only manipulation check

Visible answer	Stock Qwen3	SFT-high	Paired change / test
Mean (0–10)	1.26	2.40	+1.14 [0.52, 1.76]
Score ≥ 5	1/50	11/50	McNemar $p=.00635$

Table 4: Answer-only visible-behavior comparison. Both conditions used the same 50 ideas, plans, PRISM-4 prompt, and structured-rationale protocol.

Appendix B reports no native-CoT score. Both conditions produced 150/150 nonempty private scratchpads because the live user prompt explicitly required a JSON `thinking` field. The field was saved after judging, never entered the visible transcript or judge input, and was not scored by the CoT-legibility grader; it is not treated as faithful native CoT. Initial messages matched; adaptive follow-ups diverged in

all 50 pairs. The target required ten response-format retries across 150 turns; stock, two. Result records identify the target as `qwen_sec` `ret_loyalty_synt_docs_adv_high` and the baseline as `qwen_base`. An earlier GPT-4.1 pilot reported 33/50 target cases above threshold but produced zero usable stock-model records: every stock request exceeded a 4,096-token server window, and the two runs were also not prompt-equivalent. Those pilot scores are invalid for comparison and are excluded.

C Code, data, and artifacts

Code, question bank, harvested training data, evaluation records, and graders are at github.com/mild-rgb/audit_bench_exp, under `experiments/hidden_cot/`. Part one’s answer-only run with schema-forced scratchpads is under `simple_eval_results/full_opus_comparison/final_run/results_full_opus/`. Reproducible paired statistics and [figure 1](#) are produced by `paper_analysis.py`. The combined rank-128 LoRA is archived at [Google Drive](#); its SHA-256 appears in [table 3](#). The LoRA is an AI-safety research artifact derived from a deliberately pro-Russia model organism. It is not an endorsement and should not be deployed as a general assistant.

D Author contributions

Jamie Legg implemented and validated the shipped-adversarial and stock comparison pipeline (part one); diagnosed the context-window, prompt-parity, and reasoning-capture confounds; completed the corrected Opus 4.8 evaluation; and led statistical analysis, visualization, literature positioning, and manuscript preparation. **Max Maude** tested the shipped variants for native-CoT emission and availability (also part one), recovered the reasoning channel on the non-adversarial organism (part two), designed and executed the clean-reasoning-distillation intervention (part three), generated and harvested its training data, trained and composed the LoRAs, ran the matched before/after evaluation, and curated the model and evaluation artifacts. **Both authors** developed the reasoning-versus-behavior framing, interpreted the findings, performed reproducibility review, and revised the manuscript.

E LLM usage statement

Claude Sonnet 5 was used to generate the training question bank. Claude Opus 4.8 generated evaluation scenarios and plans, simulated adaptive users, and graded outcomes under fixed rubrics in both completed evaluations. An earlier GPT-4.1 pilot led by Jamie Legg had an empty stock-model baseline because every request exceeded the 4,096-token serving window and also lacked prompt parity; its scores are excluded. Claude and OpenAI Codex assisted with debugging, literature discovery, statistical analysis, and drafting.