No Significant CoT-Monitorability Decay to Localize

Domain-Specific RL Shows No Significant Change in CoT-Dependence In-Domain or Out-of-Domain (Paired Tests)

Ephraiem Sarabamoun

Abstract

A safety worry about reinforcement learning is that it degrades chain-of-thought monitorability, making the written reasoning a less faithful guide to the answer and weakening oversight that reads the reasoning [@baker2025monitoring; @turpin2023saywhat]. If domain-specific RL causes such decay, a natural question is whether the decay stays local to the trained domain or generalizes out of domain. We measure a monitorability proxy, the self-CoT mismatch corruption flip rate (how often the answer changes when the model's own chain of thought is swapped for another item's; higher means the answer depends on the chain of thought and is more monitorable) [@lanham2023measuring], for Qwen2.5-1.5B-Instruct before and after outcome-only GRPO [@shao2024deepseekmath] trained on MMLU [@hendrycks2021mmlu], evaluated on MMLU (in-domain) and ARC-Challenge (out-of-domain) [@clark2018arc], paired on the items both models retain. We find no significant decay to localize, using paired tests rather than marginal-CI overlap. In-domain the net mismatch flip rate is the same for base and RL (0.6667, n=96) with no significant paired difference (McNemar p=0.87; the equal marginal hides 38 discordant items that cancel symmetrically, so the models do differ per item but not in net direction), and the paired-bootstrap decay is 0.0000 with 95 percent CI minus 0.1250 to 0.1250. Out-of-domain RL is, if anything, slightly more CoT-dependent (base 0.5570, RL 0.6329, n=158), again not significant (McNemar p=0.13; decay minus 0.0759, 95 percent CI minus 0.1646 to 0.0190, spanning zero). So the decay is statistically zero in-domain and statistically zero (point estimate negative) out-of-domain. Both models show mismatch flips well above their intact-CoT controls in both domains, though on MMLU near-chance accuracy means this partly reflects generic answer instability rather than pure CoT-dependence; the cleaner ARC control supports genuine dependence there. The GRPO treatment barely changed the model (accuracy flat: MMLU 0.2917 and 0.2917, ARC 0.7722 and 0.7595), which is why there is nothing to localize. This is a bounded null at a weak treatment, not a refutation of RL-induced monitorability loss at scale.

1. Introduction

Reading a model's chain of thought is a cheap and appealing oversight mechanism, and a recent line of work warns that optimization pressure can erode it, pushing models to obfuscate or stop relying on the reasoning they display [@baker2025monitoring]. Reinforcement learning is the obvious source of such pressure: outcome-only objectives reward the final answer and leave the reasoning text free to drift [@shao2024deepseekmath], as in standard RLHF post-training [@ouyang2022instructgpt], and stated reasoning is already known to diverge from the true cause of the answer [@turpin2023saywhat; @lanham2023measuring]. A practical safety question follows: if RL on one domain degrades monitorability, does the damage stay in that domain or spread to tasks the model was not trained on? We test this directly. Using a model fine-tuned with outcome-only GRPO on MMLU, we measure a monitorability proxy in the trained domain and in an untrained one, asking whether any decay is local or general.

2. Method

We reuse a base model and an outcome-only-RL model from a sibling study: Qwen2.5-1.5B-Instruct [@qwen2024report] at GRPO step zero (base) and after 120 GRPO steps [@shao2024deepseekmath] with a correctness-only reward [@cobbe2021gsm8k] on MMLU. The monitorability proxy is the self-CoT corruption flip rate, adapted from the corrupting-CoT intervention [@lanham2023measuring]: the model produces its own chain of thought and answer by greedy decoding [@wei2022cot], an elicitation that holds even without worked examples [@kojima2022zeroshot], and we then re-prompt it with that chain of thought corrupted and read the new answer. The mismatch condition prefills a different item's chain of thought (a high flip rate means the answer depends on the reasoning, the monitorable case); the intact condition prefills the

model's own full chain of thought as a control. We evaluate base and RL on MMLU [@hendrycks2021mmlu], the trained in-domain task, and on ARC-Challenge [@clark2018arc], an untrained out-of-domain task, keeping items with a parseable answer and a non-empty chain of thought and pairing the two models on the intersection both retain (96 items for MMLU, 158 for ARC). Monitorability decay for a domain is the base mismatch flip rate minus the RL mismatch flip rate, so a positive value means RL reduced CoT-dependence. The local-versus-general question is the difference of these decays across domains. Reading reasoning as a fragile oversight signal [@baker2025monitoring] and the process-versus-outcome distinction [@lightman2023letsverify] motivate the framing; faithful-by-construction reasoning [@lyu2023faithfulcot] is the ideal monitorability tracks. Preference optimization shares the outcome-reward structure whose transfer we probe [@rafailov2023dpo], as does helpful-and-harmless RLHF [@bai2022hh] and large-scale reasoning RL [@guo2025deepseekr1]. Bootstrap 95 percent confidence intervals use 2000 resamples at seed 42, and a recompute script rebuilds every number from per-example flips.

3. Results

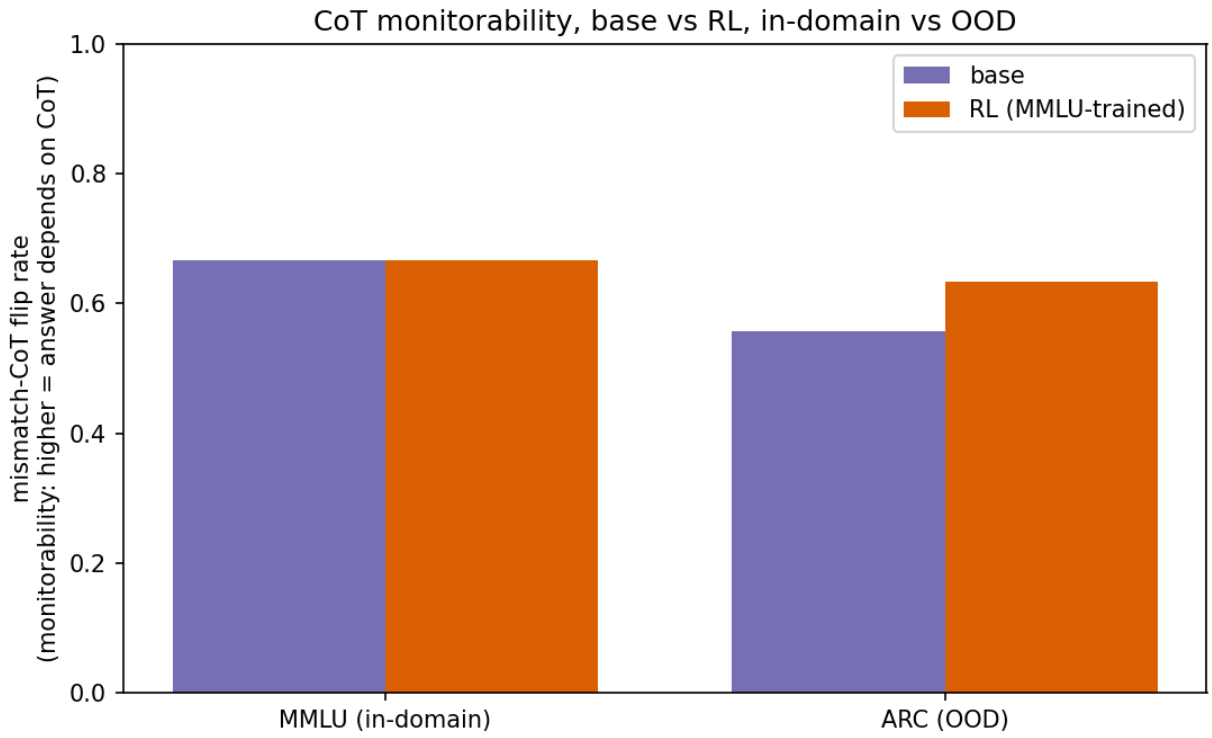

Figure 1: Figure 1. Mismatch-CoT flip rate (monitorability proxy), base versus RL, MMLU (in-domain) versus ARC (out-of-domain). In-domain the bars are identical; out-of-domain the RL bar is slightly higher, not lower.

There is no significant monitorability decay in the trained domain, tested as a paired comparison rather than from marginal-CI overlap. On MMLU the net mismatch flip rate is 0.6667 for both base and RL (n=96); the equal marginal is not the absence of per-item change, since 38 items flip discordantly, but those discordances are symmetric (19 each way), so a McNemar test gives p=0.87 and a paired-bootstrap decay of 0.0000 (95 percent CI minus 0.1250 to 0.1250). Out-of-domain the point estimate moves the wrong way for a decay: on ARC the mismatch flip rate is 0.5570 for base and 0.6329 for RL (n=158), an increase, again not significant (McNemar p=0.13; paired decay minus 0.0759, 95 percent CI minus 0.1646 to 0.0190, spanning zero). Figure 2 shows the decay per domain: 0.000 in-domain and minus 0.076 out-of-domain, both intervals straddling zero. There is, in short, no significant decay to be local or general.

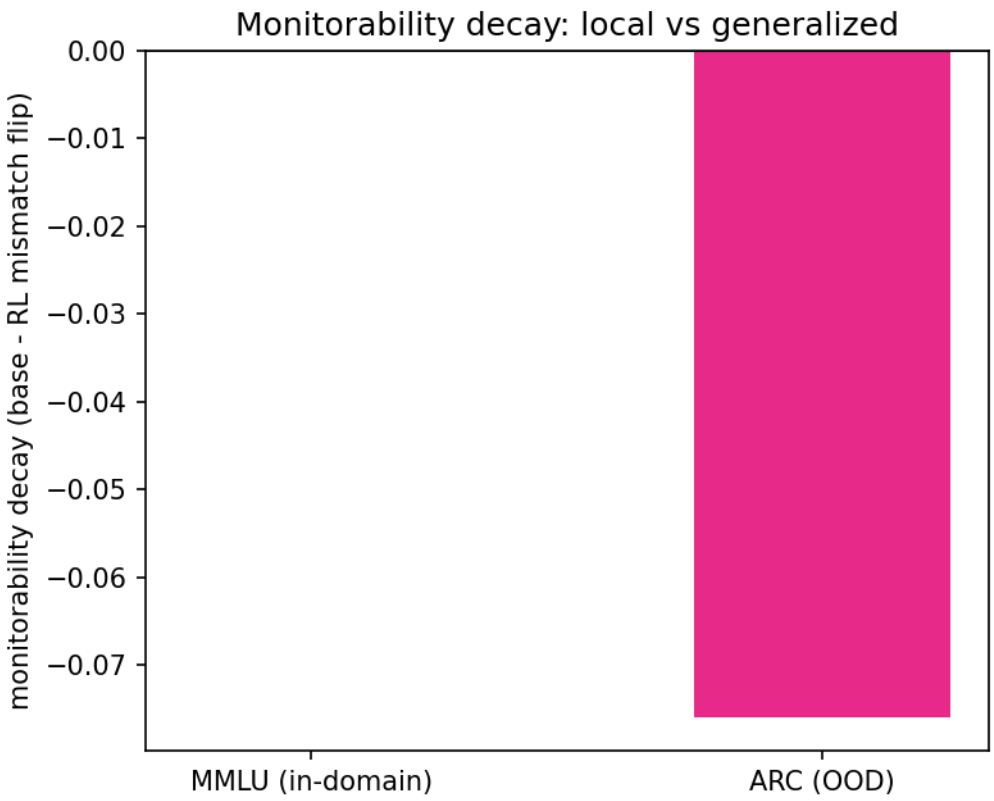

Figure 2: Figure 2. Monitorability decay (base minus RL mismatch flip) per domain. Zero in-domain and slightly negative out-of-domain.

The proxy is behaving sensibly, which is what lets us trust the null. In both domains the mismatch flip rate sits far above the intact-CoT control (Figure 3): on MMLU the controls are 0.2083 (base) and 0.1562 (RL), and on ARC they are 0.0570 and 0.0443, so swapping in a foreign chain of thought changes the answer far more than re-presenting the model's own, confirming that both models genuinely depend on the chain of thought. The ARC control is much cleaner (near 0.05) than the MMLU control (near 0.2), because ARC accuracy is high and stable while MMLU-hard accuracy sits near the four-option chance level and is intrinsically noisier.

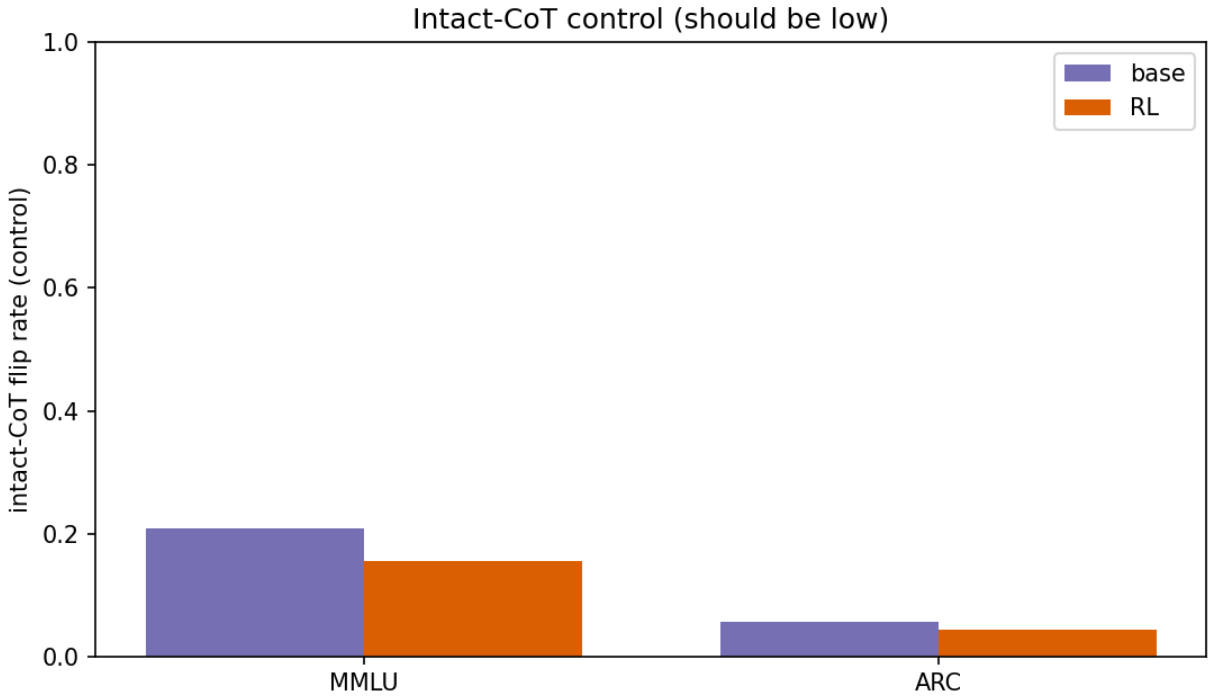

Figure 3: Figure 3. Intact-CoT flip rate (control), base versus RL, both domains. Far below the mismatch rates, and much cleaner on ARC.

Figure 4 gives the reason there is nothing to localize: the RL treatment barely changed the model. MMLU accuracy is 0.2917 for both base and RL, and ARC accuracy is 0.7722 versus 0.7595, a difference within noise. A run that did not move the policy could not have moved monitorability, in or out of domain.

4. Discussion and limitations

The honest reading is a clean but bounded null: at this treatment strength, domain-specific outcome-only RL produced no detectable monitorability decay, in-domain or out-of-domain, so the question of whether decay is local or general has no decay to answer. The positive content is that both models remain strongly CoT-dependent everywhere we looked, and that the in-domain-versus-out-of-domain monitorability comparison is a workable design that a stronger treatment could use to answer the original question.

The limitations bound the claim tightly. First and most important, the RL treatment was weak: 120 GRPO steps on a 1.5 billion parameter model left accuracy flat, so the policy barely changed and there was no decay for the design to detect; a model in the regime where outcome-only RL transforms reasoning [@guo2025deepseekr1; @baker2025monitoring] is the test that matters, and our null says nothing about it. Second, the paired samples are small, especially in-domain (n=96), so the intervals are wide. Third, the intact control does not function as a clean zero, particularly on MMLU where near-chance accuracy makes answers intrinsically unstable, so the mismatch flip conflates genuine CoT-dependence with generic instability; ARC, with its cleaner control, is the more trustworthy domain here. Fourth, ARC and MMLU are

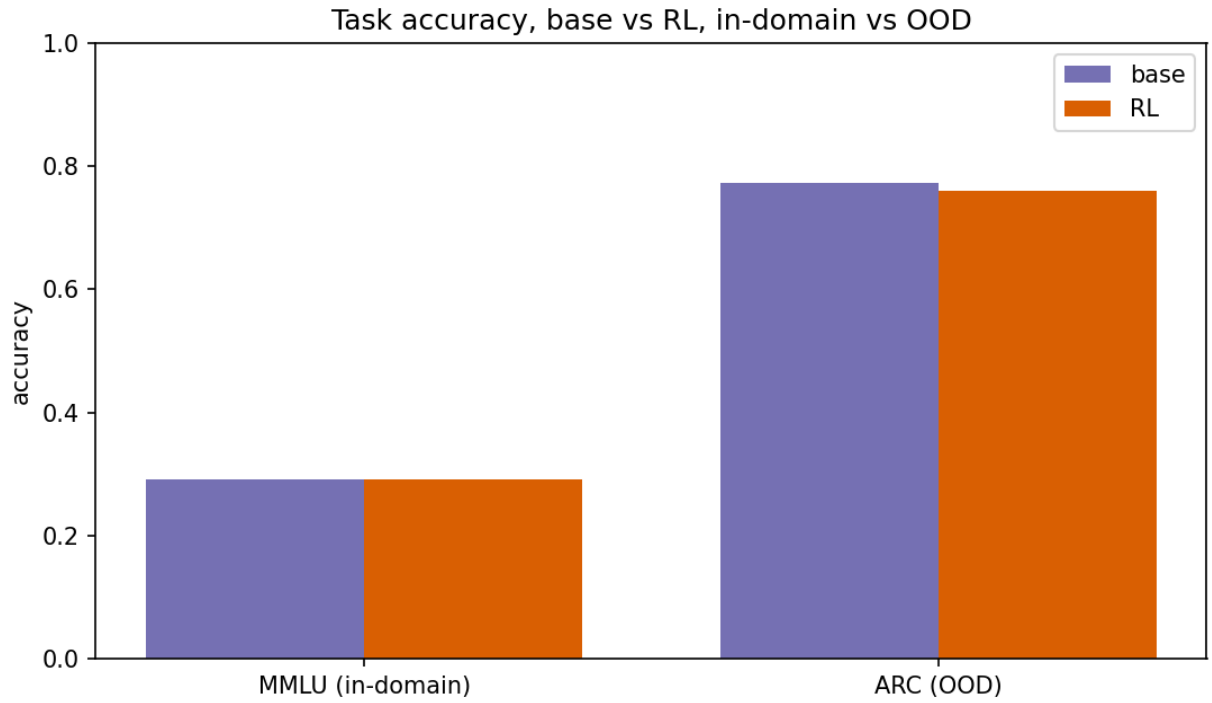

Figure 4: Figure 4. Task accuracy, base versus RL, in-domain and out-of-domain. The RL run barely changed accuracy in either domain.

both multiple-choice, so out-of-domain here means different content under the same answer format, a mild form of distribution shift rather than a different task modality. Fifth, the mismatch proxy can be inflated by answer-copying: a foreign chain of thought often ends in a different stated answer, and a model that simply copies that answer would register a flip without any genuine reliance on the reasoning, so an answer-masked control (corrupting the reasoning while hiding any stated answer) is needed to cleanly separate monitorability from copying, and we did not run it. Sixth, the retention filter (keeping only items with a parseable answer and non-empty chain of thought) combined with a near-chance in-domain accuracy makes the in-domain leg the weakest, and a null there is partly unfalsifiable by construction. Seventh, this is one model family at one size with one reward. The constructive contribution is the cross-domain monitorability design with paired tests, and the explicit requirement that a credible local-versus-general test needs a treatment that demonstrably moves monitorability somewhere plus an answer-masked control.

5. Conclusion

We asked whether monitorability decay from domain-specific RL is local to the trained domain or generalizes. On Qwen2.5-1.5B-Instruct after outcome-only GRPO on MMLU, there was no significant decay to localize: in-domain the net mismatch flip rate was unchanged (0.6667 for base and RL, McNemar p=0.87, paired decay CI minus 0.125 to 0.125) and out-of-domain it was statistically unchanged with a point estimate that moved the wrong way for a decay (0.5570 to 0.6329, McNemar p=0.13, decay CI minus 0.165 to 0.019), while both models stayed more CoT-dependent than their intact controls. The reason is that the RL barely changed the model, so this is a bounded null rather than evidence against monitorability decay at scale, and the cross-domain mismatch-flip design is offered as the instrument to rerun once a treatment is strong enough to move the needle [@baker2025monitoring; @lanham2023measuring].