Outcome-Only RL Did Not Reduce a Model's Causal Dependence on Its Own Chain of Thought

A Self-CoT Corruption-Flip Null at Small Scale

Ephraiem Sarabamoun

Abstract

A standing worry about outcome-only reinforcement learning is that rewarding only the final answer makes the chain of thought decorative, so that the answer stops causally depending on the reasoning the model writes [@lanham2023measuring; @turpin2023saywhat]. We test this directly with a self-CoT corruption-flip protocol on Qwen2.5-1.5B-Instruct before and after outcome-only GRPO [@shao2024deepseekmath] on MMLU [@hendrycks2021mmlu]. The model produces its own chain of thought and answer, and we then re-prompt it with that chain of thought corrupted in three ways, truncated to its first half, replaced by a different question's chain of thought, or removed entirely with the answer forced, and measure how often the final answer changes. An intact-CoT control, re-prompting with the model's own full chain of thought, bounds the measurement noise. The result is a null, and we are deliberately cautious about reading anything positive into it. The mismatched-CoT flip rate is statistically indistinguishable between the base model (0.6597, 95 percent CI 0.5764 to 0.7361, n=144) and the RL model (0.6543, 0.5802 to 0.7284, n=162). The no-CoT and truncate flips also differ little once their wide intervals are considered (no-CoT 0.5139 to 0.4444, truncate 0.1250 to 0.0864). We flag three confounds that keep this from being a strong null. The base and RL numbers come from different kept-item sets, so the comparison is unpaired. The truncate flip falls below the intact control (0.1250 versus 0.1944), so the intact value is not a clean lower-bound floor and the weaker-corruption comparisons are not safely interpretable as floor-relative. And accuracy is near the four-way chance level (0.3403 and 0.3272), so a high mismatch flip conflates genuine dependence on the reasoning content with generic answer instability. Above all, the RL treatment barely moved the model (accuracy 0.3403 to 0.3272), so the test had little power. We therefore report a bounded null at small scale and treat the corruption-flip-with-control protocol as a starting point that needs a paired design and an instability control before it can claim to measure causal dependence cleanly.

1. Introduction

Chain-of-thought prompting changes what models answer [@wei2022cot], an effect that appears even with no worked examples [@kojima2022zeroshot], and a natural reading treats the chain of thought as a latent that the answer is sampled from [@wang2022selfconsistency]. Whether the answer truly depends on that reasoning, or merely accompanies it, is the faithfulness question [@lanham2023measuring]. Faithful-by-construction reasoning, where the answer provably follows from the stated steps, has been proposed as the ideal [@lyu2023faithfulcot], and models are known to present explanations that are not the true cause of their answer [@turpin2023saywhat]. Outcome-only reinforcement learning sharpens the worry: methods that reward only the final answer, from RLHF [@ouyang2022instructgpt], including helpful-and-harmless training [@bai2022hh], to preference optimization [@rafailov2023dpo] to verifiable-reward RL [@shao2024deepseekmath; @guo2025deepseekr1], optimize the answer and never the reasoning, so they could in principle drive the model toward producing a correct answer while the written chain of thought becomes post-hoc decoration. Process supervision is the proposed antidote precisely because outcome supervision leaves the steps unconstrained [@lightman2023letsverify].

We ask a specific, measurable version: does outcome-only RL reduce the final answer's causal dependence on the model's own chain of thought? We adapt the corrupting-CoT intervention [@lanham2023measuring] into a before-and-after comparison on a single model and add an intact control to separate true causal dependence from measurement noise.

2. Method

We use Qwen2.5-1.5B-Instruct [@qwen2024report] and train it with GRPO [@shao2024deepseekmath] on MMLU multiple-choice questions [@hendrycks2021mmlu], with an outcome-only reward of one for the correct option letter and a small format term, never reading the chain of thought, and no answer cue in the prompt. We save a step-zero adapter, which is the base model, and a final adapter after 120 update steps. On a reserved MMLU eval slice disjoint from training, each model first produces its own chain of thought and answer by greedy decoding. We extract the answer letter and the reasoning text before the answer line, using the same parser the reward used. We then build four re-prompts that prefill the assistant turn with a version of that reasoning and read the resulting answer: intact (the model's own full chain of thought, a control that should rarely change the answer), truncate (the first half), mismatch (a different question's chain of thought from the same model), and no-CoT (the answer forced immediately with no reasoning). The flip rate for a condition is the fraction of items whose final answer differs from the model's original answer; a higher flip rate means the answer is more causally dependent on the reasoning. We keep only items with a parseable answer and a non-empty chain of thought (144 for the base model, 162 for the RL model), and report percentile bootstrap 95 percent confidence intervals with 2000 resamples at seed 42. A recompute script rebuilds every number from per-example flips. Scratchpad-style reasoning that materially drives outputs [@nye2021scratchpads] and verifiable-answer reward signals [@cobbe2021gsm8k] are the backdrop for both the prompt format and the reward.

3. Results

The intact control was intended to bound measurement noise, and it partly does, but it does not behave as cleanly as we hoped. Re-prompting each model with its own full chain of thought changes the answer 0.1944 of the time for the base model and 0.1420 for the RL model. This is well above zero, because reconstructing the assistant turn from the extracted reasoning is a soft prefill rather than a faithful replay of the original decode. We had planned to subtract this floor from the corruption flips, but as the next paragraph shows, one corruption (truncate) flips below it, so the intact value is not a valid lower bound and we do not use it as a subtraction baseline.

The mismatch condition, where the model must answer after reading a different question's reasoning, gives the largest flips: 0.6597 (95 percent CI 0.5764 to 0.7361) for the base model and 0.6543 (0.5802 to 0.7284) for the RL model. The confidence intervals almost entirely overlap, so there is no detectable base-versus-RL difference. We resist the tempting reading that this 0.66 simply means the answer depends on the reasoning, because at an accuracy of 0.34, barely above the four-way chance level of 0.25, the answer is unstable to many perturbations, and swapping in a foreign chain of thought is partly just such a perturbation; the mismatch flip therefore conflates dependence on reasoning content with generic instability. The no-CoT condition flips 0.5139 for the base model and 0.4444 for the RL model, and the truncate condition flips 0.1250 and 0.0864. We had intended to read these against the intact control as a noise floor, but that logic does not hold here: the truncate flip (0.1250) is below the intact flip (0.1944), so the control is not a valid lower bound, and the floor-relative comparison is abandoned. Figure 2 shows the mean corruption flip rate at 0.4329 for the base model and 0.3951 for the RL model, a small unpaired difference we do not interpret as a real reduction.

The most important context for the null is that the RL treatment barely changed the model. Own-CoT MMLU accuracy is 0.3403 for the base model and 0.3272 for the RL model, a difference within noise and in the wrong direction for a successful accuracy-improving RL run. Figure 3 shows this. One hundred twenty GRPO steps on a 1.5 billion parameter model on hard MMLU subjects produced almost no behavioral change, so the experiment had little opportunity to observe a change in chain-of-thought reliance even if outcome-only RL induces one at scale.

Figure 4 plots all four conditions for both models and makes the shared structure visible: the corruption severity ordering is the same for base and RL, and the two curves run close together throughout.

4. Discussion and limitations

The honest reading is a bounded, confounded null at small scale, and we list the confounds plainly because they are severe enough that we make no positive methodological claim. First and most important, the RL

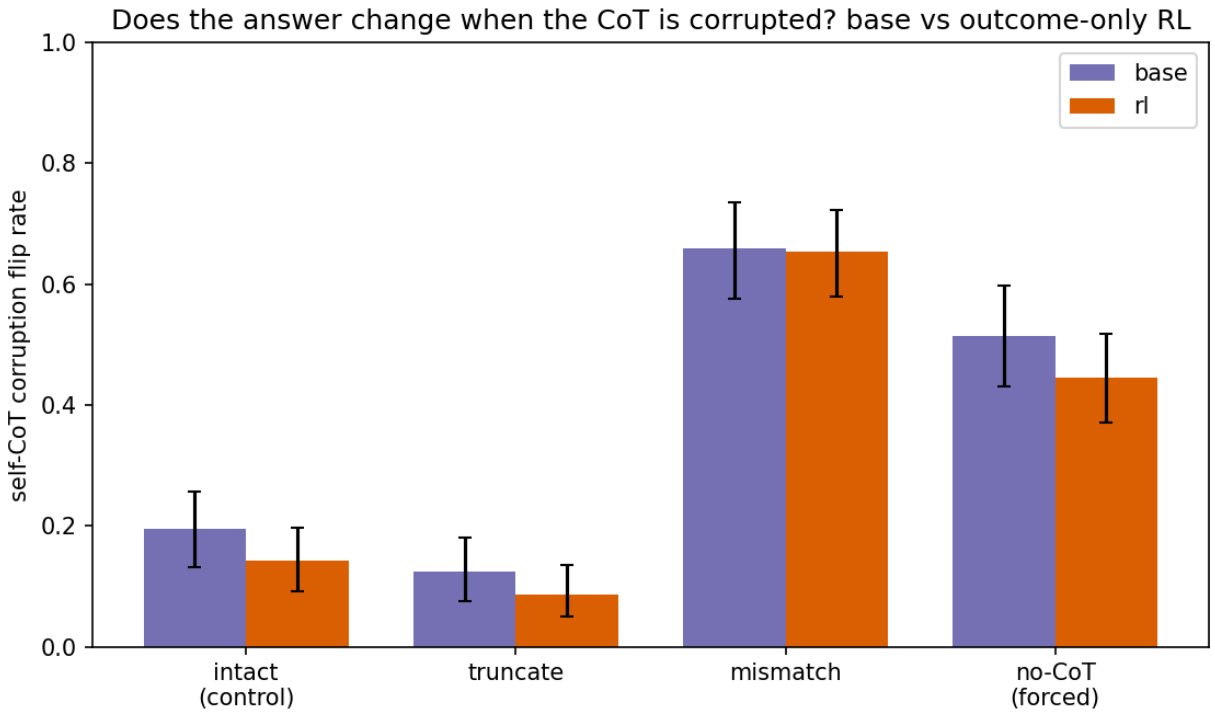

Figure 1: Figure 1. Self-CoT corruption flip rate by condition, base versus outcome-only RL, with bootstrap 95 percent confidence intervals. The mismatch and no-CoT corruptions flip the answer far more than the intact control, but the base and RL bars overlap in every condition.

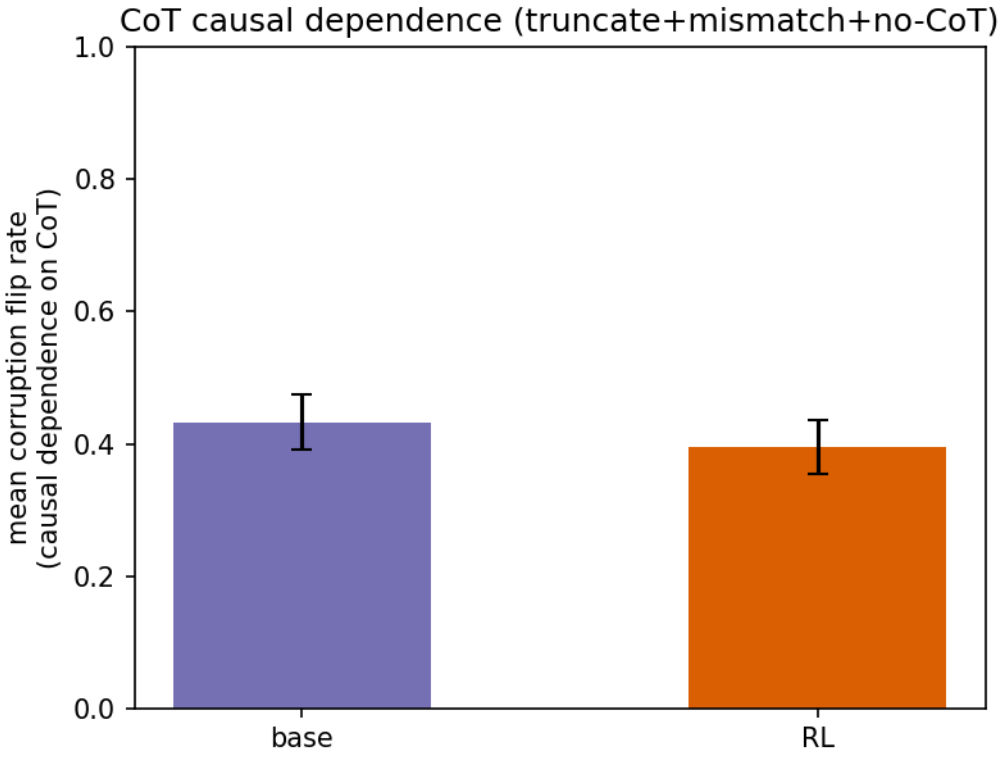

Figure 2: Figure 2. Mean corruption flip rate (truncate, mismatch, and no-CoT averaged), the summary measure of causal dependence on the chain of thought, base versus RL.

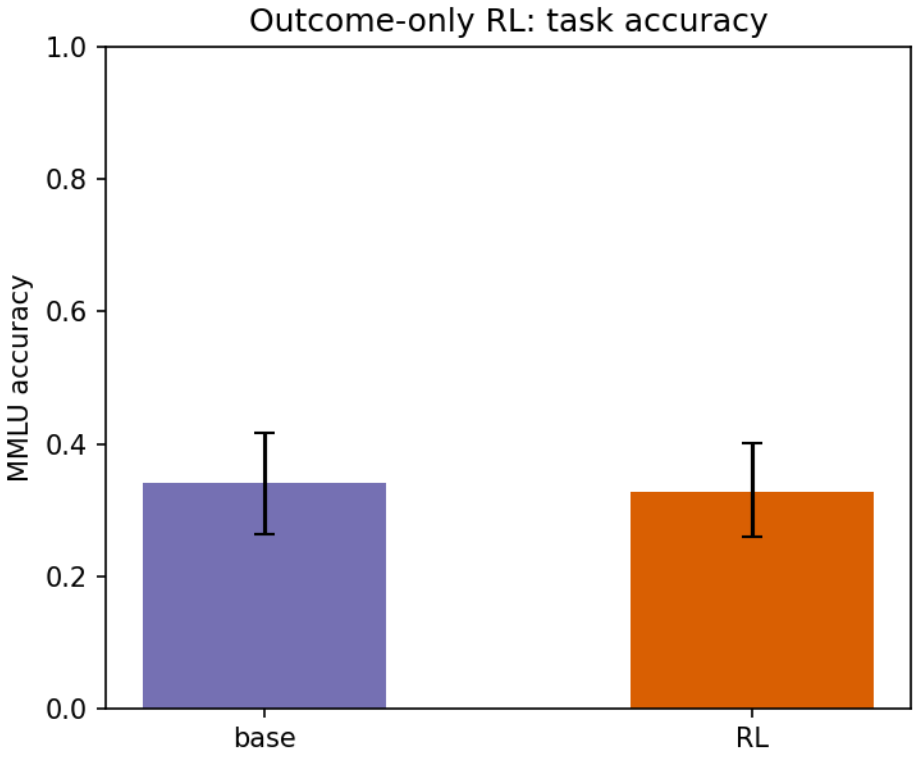

Figure 3: Figure 3. Own-CoT MMLU accuracy, base versus RL. The RL run did not improve accuracy, indicating a weak treatment.

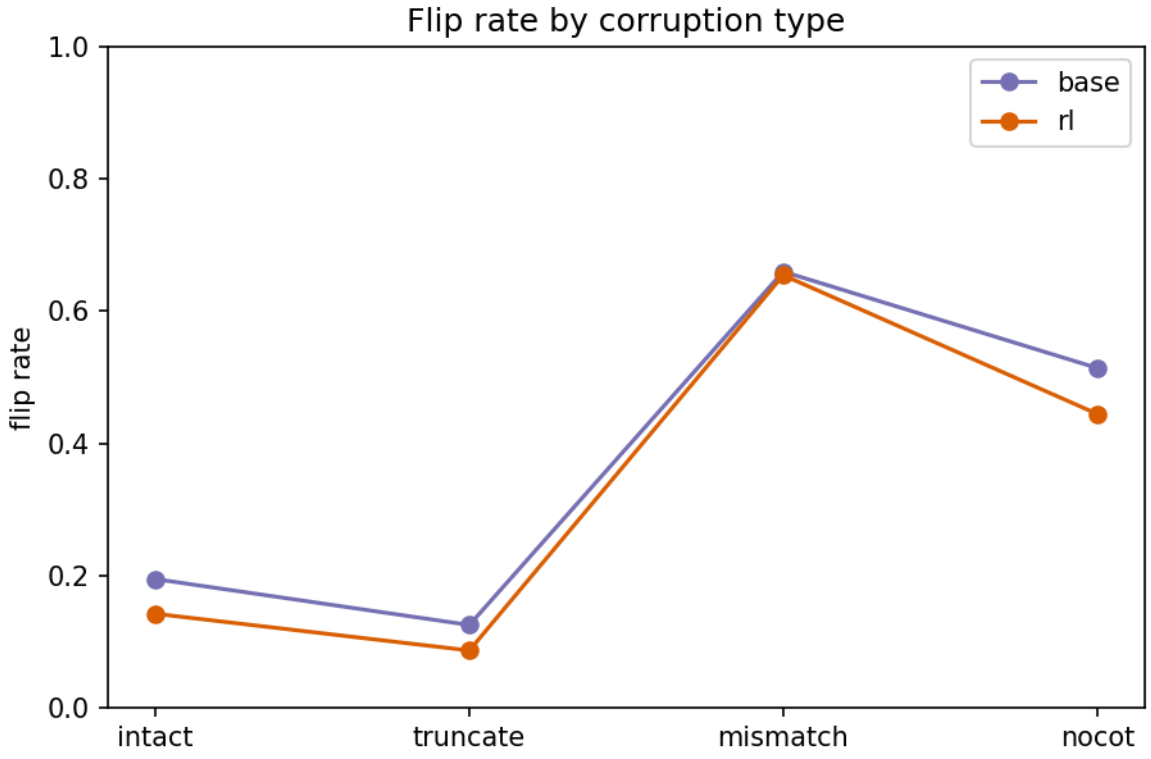

Figure 4: Figure 4. Flip rate by corruption type for both models, showing the ordering (mismatch above no-CoT above intact above truncate) is shared and the two models track each other.

treatment was weak: 120 GRPO steps on a 1.5 billion parameter model did not move accuracy, so the policy barely changed and the test had little power to detect a change in reliance. A run that demonstrably reshapes behavior, or a larger model in the regime where outcome-only RL is known to transform reasoning [@guo2025deepseekr1], is the higher-power setting. Second, the comparison is unpaired: each model is scored on its own kept-item set (144 versus 162), because which items yield a parseable answer and a non-empty chain of thought differs between the two policies, so a paired reanalysis on the item intersection is needed before any base-versus-RL difference can be trusted. Third, the intact control does not function as a clean floor: the truncate flip falls below it, which means the soft-prefill reconstruction adds idiosyncratic noise and the floor-subtraction logic we initially planned is invalid. Fourth, base accuracy is near the four-way chance level (0.34), so the answer is intrinsically unstable and a high mismatch flip cannot be cleanly attributed to dependence on reasoning content rather than to generic instability; a within-model regenerate-from-scratch instability baseline is the missing control. Fifth, MMLU multiple choice with a small model yields short, sometimes shallow chains of thought, and we measured one model family at one size with one reward. Taken together these mean the protocol, as run here, does not yet cleanly measure causal dependence; the constructive contribution is a concrete specification of the paired design and instability control a higher-power rerun would need.

5. Conclusion

We asked whether outcome-only reinforcement learning reduces a model's causal dependence on its own chain of thought, measured by how often corrupting that chain of thought flips the final answer. On Qwen2.5-1.5B-Instruct before and after 120 steps of outcome-only GRPO, we found no detectable difference: the mismatched-CoT flip rate is statistically identical across base and RL at about two thirds. We do not claim this as a clean result. The RL treatment barely changed the model, the comparison is unpaired, the intact control fails its own monotonicity check, and near-chance accuracy confounds the flip rate with generic instability, so this is a bounded and confounded null rather than a refutation of the broader concern [@lanham2023measuring]. The honest takeaway is a worked specification of what a credible test needs, a treatment that visibly reshapes the policy, a paired item set, and a regenerate-from-scratch instability control, which we set out for a higher-power rerun [@lightman2023letsverify].