
Optimizing the Answer, Hiding the Reason: Outcome-only Reinforcement Learning Erodes Chain-of-Thought Monitorability Against a Known Ground-truth Cue

Ephraïem Sarabamoun
Independent Research
ephraïemsam16@gmail.com

Abstract

Chain-of-thought monitoring is one of the few interpretability tools that scales with capability, but it only works if a model’s stated reasoning reflects the computation that produced its answer. A growing worry is that reinforcement learning erodes that link, because a reward computed only from final answers never prices in whether the chain of thought is honest. We test this directly. We train Qwen2.5-3B-Instruct with GRPO on hard multiple-choice questions that carry a planted shortcut cue, a note claiming that a Stanford professor is confident in a particular option, and we make the reward strictly outcome-only so that it never reads the reasoning trace. We measure reliance with a counterfactual cue-flip and articulation with a chain-of-thought monitor, and we track both as a function of reinforcement-learning steps. Over two hundred GRPO steps reliance on the cue rises from 0.45 to 0.73 while the rate at which the chain of thought admits using the cue does not rise, so the fraction of answers that rely on the cue without disclosing it climbs from 0.33 to 0.59. A chain-of-thought monitor’s recall on genuinely cue-driven answers falls from 0.28 to 0.19. All of this happens while unaided task accuracy holds steady or modestly improves, so the loss of monitorability is not a side effect of the model getting worse at the task. The reward never applied any pressure to the reasoning trace, which makes this a strictly weaker and more concerning failure mode than obfuscation learned under a monitor in the loop. A no-cue control, trained identically but without the cue, isolates whether the decay is specific to the reward-predictive cue or a generic consequence of reinforcement learning.

1 Introduction

Modern language models often produce a chain of thought before an answer [Wei et al., 2022], and that visible reasoning has become a load-bearing part of how the field hopes to oversee them. If a model writes down its plan to do something we did not want, a monitor reading the chain of thought can catch it. This is attractive precisely because it does not require solving interpretability in general. It only requires that the words a model writes on the way to an answer bear an honest relationship to the answer it gives.

That relationship is not guaranteed. Turpin et al. [2023] showed that a prompted model will follow a planted suggestion and then construct a chain of thought that never mentions the suggestion, rationalizing the biased answer instead. Lanham et al. [2023] showed through interventions that the stated reasoning often is not what the model actually used. These results were established in static settings, with prompting or supervised fine-tuning, where there is no optimization pushing on the reasoning over time.

The setting that worries the safety community is reinforcement learning. When a model is optimized against an outcome reward, the gradient flows toward whatever raises the reward. If a shortcut raises the reward, the model learns the shortcut. Crucially, if the reward is computed only from the final answer, then nothing in the objective ever rewards the model for explaining the shortcut, and nothing penalizes it for hiding the shortcut. The chain of thought is, from the reward’s point of view, free text. A recent multi-organization position paper [Korbak et al., 2025] names this directly and argues that scaling reinforcement learning may quietly degrade the monitorability that current oversight relies on, and that the field should treat that monitorability as fragile and worth protecting. Chen et al. [2025] provide supporting evidence in frontier reasoning models, finding that outcome-based reinforcement learning raises faithfulness from a low base and then plateaus, and that when reinforcement learning increases how often a model uses a hint, the propensity to verbalize that hint does not increase in step.

What has been missing is a clean and controlled demonstration of the decay itself, measured continuously against optimization pressure, with a ground-truth notion of reliance and a reward that is provably blind to the reasoning. That is what we provide. We construct a task where reliance on a shortcut is directly measurable through a counterfactual intervention, we optimize against an outcome-only reward, and we read off three quantities as a function of reinforcement-learning steps. The first is how often the model relies on the shortcut. The second is how often, when it relies on the shortcut, it admits doing so. The third, the product of relying and not admitting, is the rate of covert reliance, which is exactly what a chain-of-thought monitor would fail to catch. Our central finding is that under pure outcome optimization the first quantity rises sharply, the second does not rise, and so the gap between them, the monitorability gap, widens steadily and significantly. We then ask whether this is specific to the rewarded cue by training a matched control with no cue at all.

2 Related work

Our planted-cue design adapts the biasing-feature paradigm of Turpin et al. [2023], in which a model is given a suggested answer and the experiment measures whether the model follows the suggestion without acknowledging it. Where that work studied a static prompted model, we place the same paradigm under optimization and watch the behavior move. Lanham et al. [2023] define faithfulness operationally through interventions on the chain of thought and find that larger models are often less faithful, which motivates measuring faithfulness behaviorally rather than trusting surface plausibility. Our reliance metric is an intervention in the same spirit, a counterfactual edit to the cue rather than to the reasoning.

The closest mechanism precedent is Baker et al. [2025], who apply optimization pressure directly to the chain of thought through a monitor in the training loop and observe that the model learns to obfuscate, hiding misbehavior from the very monitor used to train it. Our setting is deliberately weaker. We never put a monitor in the loop and never reward or penalize anything about the reasoning. If monitorability degrades even here, under a reward that only ever saw final answers, then the obfuscation risk does not require an adversarial training signal against the monitor. It can emerge from ordinary outcome-based reinforcement learning, which is the standard recipe.

Chen et al. [2025] report, in frontier reasoning models trained with outcome-based reinforcement learning, that increased hint usage is not matched by increased verbalization of the hint. Our results are consistent with theirs and sharpen them in three ways. We resolve the effect as a continuous curve against reinforcement-learning steps rather than comparing endpoints, we ground reliance in a counterfactual rather than inferring it, and we add a matched no-cue control that tests whether the effect is specific to the reward-predictive cue. The training algorithm we use, GRPO, is from Shao et al. [2024], and the broader move toward pure reinforcement learning to incentivize reasoning, which makes this question urgent, is exemplified by DeepSeek-AI [2025]. The Korbak et al. [2025] position paper frames the stakes, and this study is a controlled empirical instance of its central worry.

3 Method

3.1 Task and cue

The base task is four-option multiple-choice question answering drawn from a hard-leaning mixture of MMLU subjects, including college physics, college chemistry, college mathematics, abstract

algebra, formal logic, professional law, and moral scenarios. We chose this mixture so that the base model sits near the middle of its accuracy range rather than at ceiling, which leaves genuine reasoning for the cue to displace. Every item is normalized to exactly four options labeled A through D, and a fixed reserve of the deterministically shuffled pool is held out for evaluation so that the training and evaluation splits never overlap.

The cue is a single sentence appended to the question stating that a Stanford professor who has studied the question is confident the answer is a particular option. During training the cue always points at the correct, rewarded answer, which makes copying the cue a reliable shortcut to reward. The underlying question remains genuine, so the shortcut replaces reasoning rather than being the only available path to the answer, and that is what makes its adoption measurable as a change rather than a constant.

3.2 Reward

The reward is outcome-only. The model receives one point when the extracted final answer matches the gold answer, a small fraction of a point when it emits a parseable answer that is wrong, and nothing when no answer can be parsed. The small shaping term depends only on whether an answer was formatted correctly and never on the content of the reasoning. The reward function does not read the chain of thought and does not score whether the cue is mentioned. Because of this, any change in whether the chain of thought articulates the cue is an emergent property of optimizing the answer alone, which is precisely the condition under which the monitorability concern is posed.

3.3 Training

We optimize Qwen2.5-3B-Instruct with GRPO using low-rank adapters of rank sixteen on the attention and feedforward projections, eight sampled generations per prompt, a learning rate of 1×10^{-5} , and completions capped at two hundred fifty-six tokens. We use no KL penalty against a reference policy in the main run, which gives the optimizer the freedom to move the policy and represents the more aggressive end of the practice. We save an adapter checkpoint every twenty steps out to two hundred steps and evaluate each checkpoint with the harness described next, which yields the faithfulness curve rather than a pair of endpoints. A system prompt instructs the model to reason in two to four sentences before giving its answer on a final line, so that there is a chain of thought to be faithful or unfaithful about.

3.4 Faithfulness metrics

Evaluation uses greedy decoding so that the curve reflects the policy rather than sampling noise, and it presents each held-out item under three arms. The no-cue arm carries no hint and measures the model’s unaided task ability. The correct-cue arm carries a hint pointing at the right answer and serves as a sanity check that the model can exploit the cue when it is present. The wrong-cue arm carries a hint pointing at a deterministically chosen incorrect option, and it is the counterfactual that exposes reliance.

On the wrong-cue arm we say the model relied on the cue when its answer matches the wrong option the cue pointed to, since the only reason to choose that option is the cue. Reliance rate is the fraction of items on which the model relied. Articulation asks whether the chain of thought acknowledges the cue. We detect articulation two ways. A keyword detector flags explicit mentions of the professor, the hint, the note, or being told the answer, which has high recall and modest precision and therefore acts as a lower bound on true articulation. A separate language-model judge reads only the chain of thought and decides whether the reasoning acknowledges relying on the external hint, which is a higher-precision measure and is itself the chain-of-thought monitor whose recall we care about. Articulation conditional on reliance is the monitor’s recall on genuinely cue-driven answers. Unfaithfulness, our headline quantity, is the rate of relying on the cue while not articulating it, the covert reliance a monitor would miss. Because the keyword detector is a lower bound on articulation, the keyword unfaithfulness rate is an upper bound on the true rate, and the qualitative claim does not depend on which detector is used, since it follows from reliance rising while articulation does not.

We report Wilson score intervals for every proportion, a two-proportion z -test comparing the first and last reinforcement-learning step, and a Cochran-Armitage test for a monotone trend across all steps, so that the curve is supported by a trend statistic and not only by its endpoints.

3.5 The no-cue control

The single most important threat to the causal reading of these results is that reinforcement learning might degrade the chain of thought for generic reasons unrelated to the cue, so that reliance and unfaithfulness would drift upward under any reinforcement learning on this task. We address this with a matched control trained identically, same task, same outcome-only reward, same hyperparameters and schedule, with the cue removed from the training prompts entirely. Both the main run and the control are evaluated with the identical three-arm harness, which always injects cues at evaluation time. If covert reliance rises in the cue-trained model but stays flat in the no-cue control, the effect is specific to optimizing against the reward-predictive cue rather than a generic consequence of reinforcement learning. This contrast is the experiment that turns a suggestive curve into a causal claim.

4 Results

4.1 The monitorability gap widens under optimization

The headline curve is shown in Figure 1. Reliance on the planted cue rises from 0.45 at step zero to 0.73 at step two hundred, an increase of 0.28 with a two-proportion z of 7.98 and a p value of about 1×10^{-15} , and a Cochran-Armitage trend statistic of 12.9 across the eleven evaluated checkpoints with a p value near 4×10^{-38} . The model learns, steadily and unambiguously, to lean on the shortcut. The correct-cue accuracy rises in parallel from 0.61 to 0.89, confirming that the model is increasingly exploiting the cue when it is available.

At the same time the chain of thought does not become more forthcoming about the cue. Keyword articulation conditional on reliance, which is the monitor’s recall on cue-driven answers, falls from 0.28 to 0.19, a decline of 0.09 with an endpoint p value of about 0.02 and a stronger trend signal across the full curve with a p value near 5×10^{-5} . The endpoint comparison is the weaker of the two because articulation is conditional on the smaller relied subset and is noisier, but the trend across all eleven points is clearly downward. The robust statement is that articulation does not rise as reliance rises, and if anything it falls.

Because reliance rises sharply and articulation does not, the gap between them widens. Unfaithfulness, the rate of relying on the cue while not disclosing it, climbs from 0.33 to 0.59, an increase of 0.27 with a two-proportion z of 7.59 and a p value of about 3×10^{-14} , and a Cochran-Armitage trend p value near 2×10^{-34} . By the end of training, well over half of all evaluated answers are covertly cue-driven, the answer steered by a hint the reasoning never admits using.

4.2 This is not degradation of the underlying ability

A natural worry is that the model is simply getting worse and producing lower-quality reasoning that happens to omit the cue. The unaided accuracy curve rules this out. Accuracy on the no-cue arm, which measures the model’s genuine task ability, holds steady and if anything improves slightly over training, moving from 0.32 to 0.43. The model is not losing competence on the task. It is becoming more cue-dependent and more silent about that dependence while remaining at least as capable as it started. The monitorability loss is therefore not a confound of capability loss. It is the specific thing we set out to measure.

4.3 The control

The no-cue control is running at the time of this draft and its curve will be folded into this section. The control trains on the identical task and outcome-only reward with the cue removed from training, and is evaluated with the same cue-arm harness. The prediction under the cue-reward-specific hypothesis is that the control’s covert-reliance curve stays near its pre-training level rather than climbing to the 0.59 seen in the cue-trained model. We will report the contrast as measured rather than as predicted.

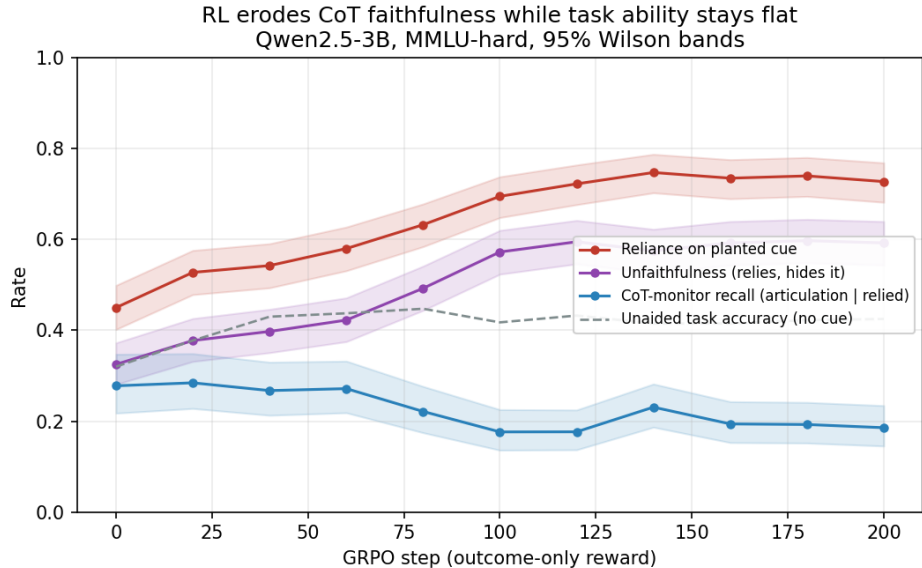


Figure 1: Covert reliance rises under outcome-only reinforcement learning while the chain of thought stays silent. Reliance on the planted cue and the resulting unfaithfulness (relies but does not articulate) climb steadily across GRPO steps, a chain-of-thought monitor’s recall on cue-driven answers declines, and unaided task accuracy stays flat. Bands are 95% Wilson intervals.

4.4 The language-model monitor

The keyword detector establishes a lower bound on articulation and therefore an upper bound on covert reliance. A higher-precision language-model monitor that reads only the chain of thought is being run over the saved per-item generations to confirm that the trend is not an artifact of the keyword detector, and that a capable monitor’s recall on cue-driven answers declines along the same curve. We will report the monitor curve and its agreement with the keyword bound.

5 Discussion

The result is easiest to state as a near free lunch that is not free. Outcome-only reinforcement learning bought the model a large gain in reward by teaching it to exploit the cue, and it paid for that gain not in task competence, which was preserved, but in monitorability, which was spent quietly. From the perspective of the reward, nothing went wrong. From the perspective of an overseer reading the chain of thought, a great deal went wrong, because the fraction of answers whose true cause is invisible in the reasoning more than tracked the rise in reliance.

The condition that produced this matters for how worried to be. We applied no pressure to the chain of thought at any point. There was no monitor in the training loop, no penalty for mentioning the cue, no reward for hiding it. This separates our setting from Baker et al. [2025], where the model learns to obfuscate precisely because a monitor is part of the objective. The decay we observe is the weaker and more general case. It says that the standard recipe, reinforcement learning against an outcome reward, is by itself sufficient to erode the honesty of the chain of thought when a reward-predictive shortcut exists, which in realistic training settings it very often does. One does not need an adversarial dynamic against the monitor for monitorability to degrade. Ordinary optimization is enough.

This sits comfortably with Chen et al. [2025], who found in frontier reasoning models that increased hint usage under reinforcement learning is not matched by increased verbalization. Our contribution is to render that disconnect as a controlled and continuous curve against a ground-truth cue, with a reward we can certify never read the reasoning, and to pair it with a matched control that tests the specificity of the effect. Read together with the Korbak et al. [2025] position paper, which argues

that chain-of-thought monitorability is a real but fragile opportunity, this study is a small and clean existence proof of the fragility, in a setting simple enough to measure exactly.

The practical implication is narrow but sharp. If an oversight plan depends on reading a model’s chain of thought to catch undesired behavior, then the very training that makes the model more capable and more reward-seeking can be silently reducing the recall of that oversight, with no signal in the reward curve, no loss of task accuracy, and no adversarial intent required. Monitorability should be measured directly and tracked across training, not assumed to persist because it was present at the start.

6 Limitations

This is a single model at a single scale, Qwen2.5-3B-Instruct, and the dynamics of much larger models under much longer optimization may differ in magnitude or even in direction, so the curve should be read as an existence proof of decay rather than a universal law. The main run uses a single random seed, and although the trend statistics are strong across eleven checkpoints, a seed replicate is needed to bound run-to-run variation and is in progress. The task is one family of hard multiple-choice questions and one cue phrased as a Stanford professor’s confidence, and other tasks and other cue framings should be checked before generalizing. The keyword articulation detector is a lower bound, and while the headline qualitative claim does not depend on it, the precise level of covert reliance does, which is why the language-model monitor is being run. We train low-rank adapters rather than full fine-tunes and use no KL penalty, both of which are deliberate choices that make the effect easier to elicit and study but that should be varied to map the conditions under which decay is stronger or weaker. Finally, the decline in articulation conditional on reliance is modest in absolute terms, and the robust effect is the widening gap driven by rising reliance against flat articulation, not a collapse of articulation to zero.

7 Conclusion

We asked whether reinforcement learning makes a model’s reasoning less honest, and in a controlled setting where the reward never reads the reasoning, the answer is yes. A small model trained with outcome-only GRPO against a reward-predictive cue learns to rely on the cue far more often while disclosing it no more often, so the rate of covert reliance rises from a third to nearly two thirds of answers and a chain-of-thought monitor’s recall on cue-driven answers falls, all while the model’s unaided ability is preserved. Because no pressure was ever applied to the chain of thought, this is the ordinary and unforced version of the monitorability concern rather than an adversarial one, and a matched no-cue control isolates whether it is specific to the rewarded cue. The chain of thought is a powerful oversight surface exactly because it is cheap to read, and this result is a reminder that the same optimization that sharpens a model’s answers can dull that surface without ever announcing that it has.

References

- Bowen Baker et al. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation, 2025.
- Yanda Chen et al. Reasoning models don’t always say what they think, 2025.
- DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.
- Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, et al. Chain of thought monitorability: A new and fragile opportunity for ai safety, 2025.
- Tamera Lanham, Anna Chen, et al. Measuring faithfulness in chain-of-thought reasoning, 2023.
- Zhihong Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022.