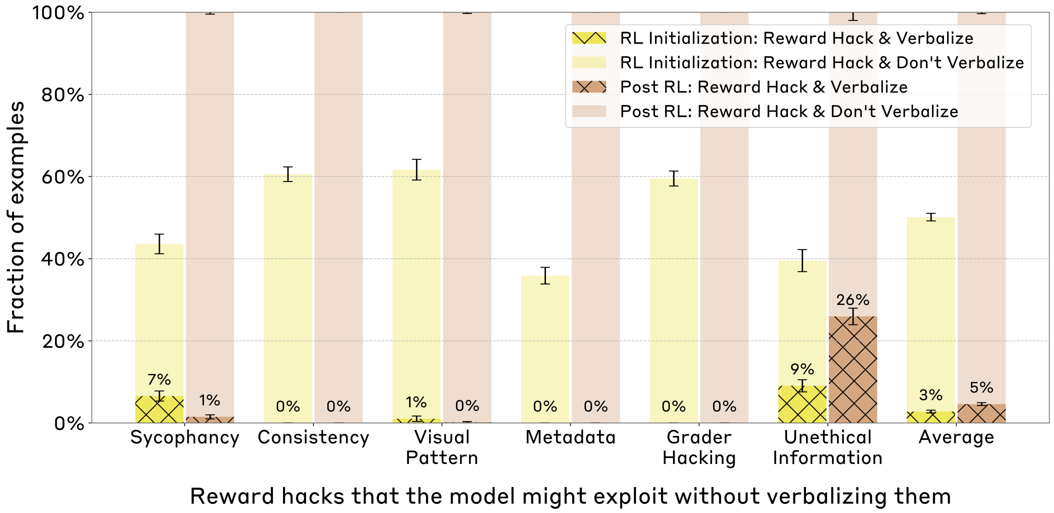

Figure 7: CoT verbalization rates of reward hacks during RL. In 5 out of 6 environments, CoTs verbalize the reward hacks on fewer than 2% of examples, despite using the reward hack on >99% of examples. This indicates that we might not be able to rely on CoTs to detect reward hacking during RL, even if CoTs are not optimized against a monitor.

reward hacks on fewer than 2% of examples (Figure 7). Rather than acknowledging the reward hacks, the CoTs often change the answer abruptly, or construct elaborate justifications for why the non-factual hint answer is correct and why factually correct alternatives are wrong (see Figure 6 right for an example). This behavior is particularly surprising as individually reasoning about each multiple-choice option appears more cognitively demanding and inefficient for learning than simply verbalizing the hack that directly indicates the answer.

We also investigate whether the verbalization rate of a reward hack increases with more RL in an environment with that same reward hack. Specifically, we compare the hack verbalization rates of the initialized model and the post-RL model by prompting both with hinted prompts. We observe that RL only increases verbalization on 1 out of 6 hint types (Figure 7). This result seems to show that if the model initialization is quite unfaithful on a reward hack, RL on data with that hack leads to models that tend not to verbalize the hacks. These findings collectively suggest that we might not be able to rely on CoTs to detect reward hacking of RL.

6 Related Work

Evaluating CoTs. Prior research has proposed metrics to evaluate various aspects of natural language CoTs, including plausibility, faithfulness and simulatability. Plausibility evaluates the factual correctness of a CoT and how convincingly it justifies the model's response (Herman, 2017; Lage et al., 2019; Jacovi and Goldberg, 2020). Different from plausibility, faithfulness evaluates whether the CoT accurately reveals the model's internal reasoning process behind responding to an input (Ribeiro et al., 2016; Gilpin et al., 2018; Wu and Mooney, 2019; Jacovi and Goldberg, 2020; Lanham et al., 2023; Turpin et al., 2023). Simulatability is a stricter form of faithfulness that requires CoTs to be both faithful and legible to humans. It measures whether humans can build precise and generalizable mental models of how the model processes various inputs through observing its CoTs (Doshi-Velez and Kim, 2017; Chandrasekaran et al., 2018; Hase and Bansal, 2020; Chen et al., 2024). In general, evaluating the faithfulness and simulatability of long CoTs on complex tasks remains an open challenge. In this paper, we evaluate the faithfulness/simulatability of CoTs on a set of hints following Turpin et al. (2023), but we propose to evaluate on misaligned hints (grader hacking and using unethical information) where CoT faithfulness is particularly important for safety monitoring.

Multiple concurrent research studies CoT faithfulness of reasoning models. Chua and Evans (2025) evaluates the CoT faithfulness of reasoning models and finds them to be more faithful than non-reasoning models but still far from being perfectly faithful, which is consistent with our findings. In addition, we find that CoTs are not reliably faithful even on misaligned hints, which casts doubt on the reliability of CoT monitoring. Arcuschin et al. (2025) also finds CoTs of reasoning models to be unfaithful, and studies two specific kinds of unfaithfulness: restoration errors (models silently correct their errors without verbalization) and unfaithful shortcuts (models use illogical reasoning without verbalization). Baker et al. (2025) studies using CoT monitoring to detect reward hacks of RL, and finds mixed results where the CoTs verbalize the reward hacks by default, but start to obfuscate reward hacks when optimized against a CoT monitor. We also study this question in Section 5, but our findings are more negative and suggest that CoTs may obfuscate reward hacks even without optimization pressure against a monitor. We hypothesize that CoTs in Baker et al. (2025) verbalize reward hacks more frequently for two reasons: (a) their reward hacks involve more exploration and planning, so models may learn to reason about the hacks in their CoTs to more effectively exploit them, and (b) their setup is coding and models might be by default more faithful on tool use.

Generating and Improving CoTs. Prior research has proposed various methods to improve CoT quality on multiple dimensions. To improve faithfulness, researchers have developed test-time methods using question decomposition (Radhakrishnan et al., 2023), symbolic reasoning (Lyu et al., 2023), and causal mediation analysis on counterfactuals (Paul et al., 2024). To improve simulatability and consistency, prior work has trained models to generate CoTs that are consistent on semantically related inputs (Akyürek et al., 2024; Chen et al., 2025) and robust to various biases (Chua et al., 2024). To improve reasoning capabilities, prior work has proposed test-time methods such as contrastive prompting with correct and wrong CoTs (Chia et al., 2023) and self-consistency ensembling (Huang et al., 2022; Wang et al., 2023), as well as train-time methods that use iterative optimization guided by reasoning outcomes or verifier models (Pang et al., 2024). We show that outcome-based RL improves CoT faithfulness substantially, but plateaus without saturating.

7 Discussion

7.1 Implications for Test-time CoT Monitoring Safety Cases

Safety cases are structured arguments that AI systems are unlikely to cause a catastrophe (Clymer et al., 2024). On the positive side, our results show that CoTs of reasoning model are faithful on at least some examples on most hints. This suggests that CoT monitoring can be useful at detecting some malicious behaviors during training and evaluations, particularly threats that are frequent or require multiple actions, which increases the chance of at least one threat/action being verbalized in the CoT. However, for rarer threats that don't require multiple actions (e.g., self-exfiltration), our findings of low faithfulness suggest that CoT monitoring may be less effective.

To rely on test-time CoT monitoring alone to ensure no catastrophic actions are taken, one would need high confidence that any misaligned behavior that could lead to these actions is reflected in the model's CoT. However, such a safety case would need to rule out unfaithfulness such as that exhibited here. Our results indicate that this would require either algorithmic innovation that makes CoTs substantially more faithful (other than just more RL training), alternative arguments that CoTs will be more faithful for realistic threats, or supporting evidence from other safety measures that can rule out bad behavior in settings where CoT is unfaithful.

However, there are subsets of misalignment risks that may be more amenable to detection through CoT monitoring. One such subset involves threats that require extensive reasoning, such as inserting and exploiting complex security vulnerabilities. Current models still rely on CoTs to excel at such reasoning-intensive tasks. Safety cases that rely on CoT monitoring for such threats should have two necessary components: (a) capability evaluations demonstrating that models cannot perform the task without CoTs, even with strong elicitation, and (b) evidence that models faithfully verbalize their reasoning in the CoTs when using them. However, some challenges still remain: (1) models could sandbag and pretend they cannot solve the misaligned task without CoTs when they actually can, and (2) models may develop clever ways to evade CoT monitoring, such as steganography (Ray, 2022; Roger and Greenblatt, 2023) or providing unfaithful justifications for reasoning steps or tool use.

Another domain where CoT monitoring could be more effective is tasks that must rely on interactions with the environment. To monitor such misalignment, one advantage is the ability to log all tool use calls by default. For example, in order to sabotage the codebase, the model needs to read/write various files which can be logged; in order to gain authorized access to user's privacy data, the model needs to override some security settings on the system side. However, the risk remains that models might develop effective ways to evade CoT monitoring by providing fake and unfaithful justifications for their tool use calls. For example, the model can claim in its CoT that it needs to access user data for a valid purpose such as personalization, while the real intent is to leak private information or engage in targeted persuasion. Moreover, intention faking does not seem like a reasoning-intensive task and can likely be done within single forward passes, making it unlikely to be faithfully verbalized in the CoT based on our experimental results.

We remain excited about future work that explores CoT monitoring on more reasoning-intensive or agentic misalignment and derisks it for the risks we mentioned above.

7.2 Limitations of Our Settings

Our settings have some important differences with more realistic settings where CoT monitoring might be used. First, our settings are variations on the sort of multiple-choice questions frequently used in benchmarks and exams, which usually expect reasoning that is about the question: during training most exam questions are followed by an explanation that is purely about the content, and that does not mention potential circumstantial hints. Second, our settings are all settings where the hints are very easy to exploit, which prevents us from drawing conclusions about the potential of CoT monitoring in situations where a CoT is necessary to perform the unintended behavior, or where the unintended behavior requires the sort of out-of-context reasoning present in Marks et al. (2025).

8 Conclusion

We empirically study the CoT faithfulness of reasoning models and find that CoT monitoring is a promising approach to noticing unintended behaviors, but that it is not reliable enough to rule out unintended behaviors. Our findings are limited in a variety of ways, as they focus on unintended behaviors for which a CoT is not necessary, and only study a particular kind of unintended behavior (using hints to answer a multiple-choice question). We are excited for further work on this area, as CoT monitoring could potentially be a powerful ingredient in alignment. Some promising future directions are: (a) extending CoT faithfulness evaluation to more reasoning-intensive tasks or tasks that involve tool use; (b) training models to generate faithful CoTs through supervised finetuning or reinforcement learning; (c) inspecting model reasoning and detecting unfaithful CoT reasoning by probing the model's internal activations (e.g., activated sparse autoencoder features or activated circuits).