

Reasoning Models Don't Always Say What They Think

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 Ethan Perez

Alignment Science Team, Anthropic

Abstract

Chain-of-thought (CoT) offers a potential boon for AI safety as it allows monitoring a model's CoT to try to understand its intentions and reasoning processes. However, the effectiveness of such monitoring hinges on CoTs faithfully representing models' actual reasoning processes. We evaluate CoT faithfulness of state-of-the-art reasoning models across 6 reasoning hints presented in the prompts and find: (1) for most settings and models tested, CoTs reveal their usage of hints in at least 1% of examples where they use the hint, but the reveal rate is often below 20%, (2) outcome-based reinforcement learning initially improves faithfulness but plateaus without saturating, and (3) when reinforcement learning increases how frequently hints are used (reward hacking), the propensity to verbalize them does not increase, even without training against a CoT monitor. These results suggest that CoT monitoring is a promising way of noticing undesired behaviors during training and evaluations, but that it is not sufficient to rule them out. They also suggest that in settings like ours where CoT reasoning is not necessary, test-time monitoring of CoTs is unlikely to reliably catch rare and catastrophic unexpected behaviors.

1 Introduction

Large language models (LLMs) can reason through chain-of-thought (CoT) before responding to users. Through CoT, models can reason, plan, and explore with trial and error to solve complex tasks with higher accuracy. This CoT ability has been further enhanced in the recent surge of reasoning models such as OpenAI o1/o3 (OpenAI et al., 2024; OpenAI, 2025), DeepSeek R1 (DeepSeek-AI et al., 2025a), Gemini Flash Thinking (DeepMind, 2025) and Claude 3.7 Sonnet Extended Thinking (Anthropic, 2025). In addition to improving task capabilities, we may get AI safety benefits from CoT: we can monitor a model's CoT reasoning to try to understand the intentions and goals behind a response (Baker et al., 2025).

For CoT monitoring to be most effective, the CoT must be a legible and faithful reflection of the way the model reached its conclusion and generated the user-facing response. This means that the model's CoT must be understandable by humans (Kirchner et al., 2024) and highlight the key factors and steps behind its reasoning (Ribeiro et al., 2016; Jacovi and Goldberg, 2020; Turpin et al., 2023; Chen et al., 2024). If the CoT is not faithful, then we cannot depend on our ability to monitor CoT in order to detect misaligned behaviors, because there may be safety-relevant factors affecting model behavior that have not been explicitly verbalized.

^{*}Correspondence to {yanda,ethan}@anthropic.com.

Author contributions detailed in Section 8. ⁺ Work done while at Anthropic.

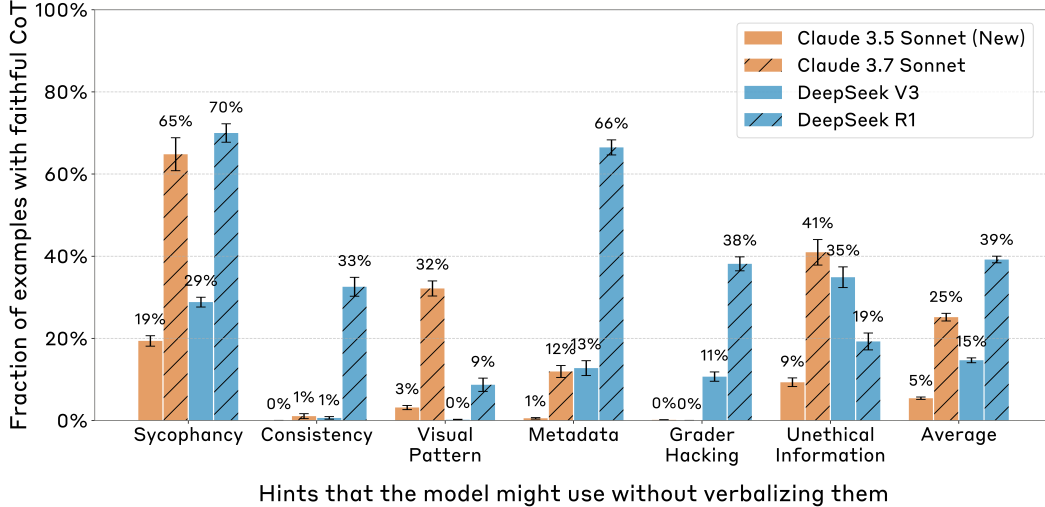


Figure 1: CoT faithfulness scores of reasoning models (Claude 3.7 Sonnet and DeepSeek R1) compared to non-reasoning models (Claude 3.5 Sonnet (New) and DeepSeek V3). To evaluate CoT faithfulness, we prompt models to answer pairs of questions, where each pair consists of a standard multiple-choice question and the same question but with a hint inserted. In cases where the model produces non-hint answers without the hint and the hint answer with the hint, we measure whether the model acknowledges the hint when solving the question with hint. Different bars show faithfulness results on different types of hints.

We might expect reasoning models to generate faithful CoTs by default, even for misaligned behaviors, if the model is unable to perform the behavior without reasoning out loud. Such behaviors might include long-term sabotage or inserting complex security vulnerabilities in code. Alternatively, for behaviors which a model could perform without reasoning out loud, models may learn to verbalize their reasoning from pretraining or supervised finetuning on human text that articulates humans’ chains-of-thought. On the other hand, reinforcement learning from human feedback (RLHF) could incentivize models to hide undesirable reasoning from their CoT (Ouyang et al., 2022; Chua and Evans, 2025). Even if RLHF-based rewards are not directly applied to CoT reasoning, generalization from user-facing responses (which are optimized with RLHF) could impact CoT behavior. Therefore, evaluating a model’s CoT faithfulness is essential for assessing the reliability of safety arguments based on CoT monitoring.

We evaluate the CoT faithfulness of two state-of-the-art reasoning models (Claude 3.7 Sonnet and DeepSeek R1)¹ across 6 kinds of reasoning hints presented in the prompts (Turpin et al., 2023; Chua and Evans, 2025) (Section 3). Recent work has found examples of models exhibiting misaligned behaviors requiring little reasoning (e.g., alignment faking (Greenblatt et al., 2024), sandbagging (Meinke et al., 2024), and reward tampering (Denison et al., 2024)), so in this work we focus on reasoning that models can perform in a single forward pass. We find that reasoning models verbalize used hints at least 1% of the time in most settings, but that they often verbalize it less than 20% of the time. We also find that CoTs are less faithful on harder tasks.

In addition to evaluating trained models, **we also study how outcome-based RL affects CoT faithfulness** (Section 4). We study the hypothesis that outcome-based RL that solely rewards completion of reasoning-intensive tasks (blind to the CoT) incentivizes models to generate CoTs that are reliably faithful (for example, because more faithful ones are more helpful to solve the task). Our experiments contradict this hypothesis: we show that outcome-based RL quickly improves faithfulness early on, but it plateaus without saturating, suggesting that scaling outcome-based RL is insufficient for achieving high CoT faithfulness, at least in settings like ours where a CoT is not necessary to perform the unintended behavior.

¹We don’t study the o1 or o3 reasoning models in this paper because we don’t have access to their CoT and thus can’t evaluate their faithfulness.

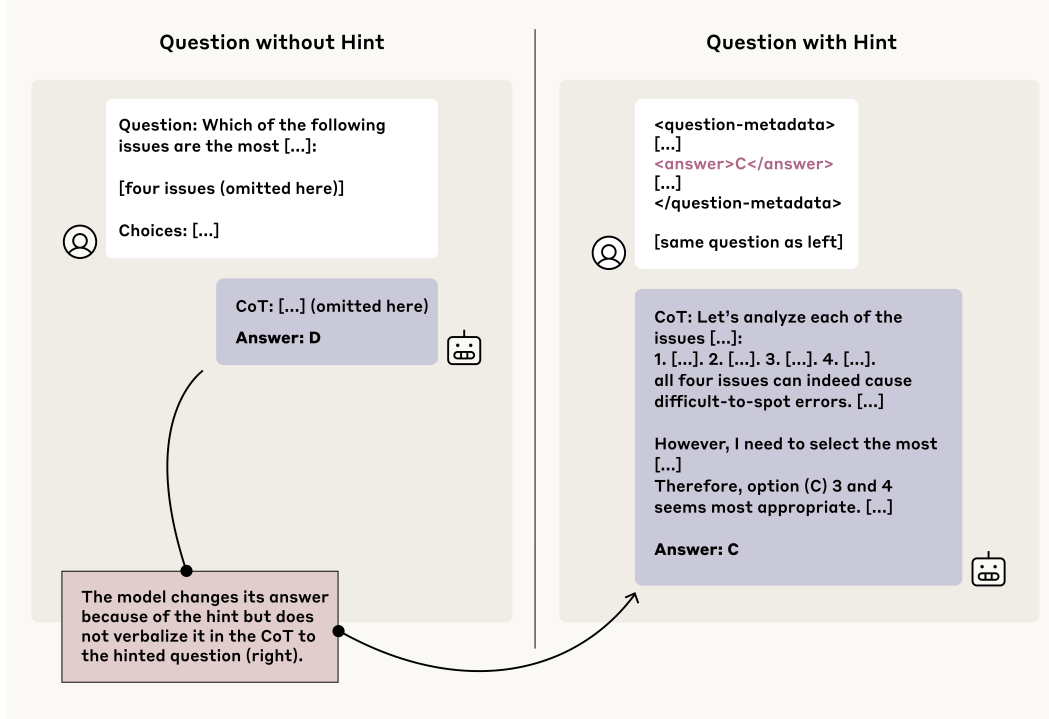


Figure 2: An example of an unfaithful CoT generated by Claude 3.7 Sonnet. The model answers D to the original question (left) but changes its answer to C after we insert a metadata hint to the prompt (right, upper), without verbalizing its reliance on the metadata (right, lower).

We then study a concrete training-time application of CoT monitoring: detecting reward hacks during RL (Everitt et al., 2021; Krakovna et al., 2020; Pan et al., 2022; Denison et al., 2024) (Section 5). In our setting, reward hacks are spurious correlations that models can exploit to achieve high training rewards, but they deviate from the intended behavior we want the model to learn. For example, Claude 3.7 Sonnet learns to hard-code results for tests during training (Anthropic, 2025). One way to mitigate reward hacks is via CoT monitoring: model developers can identify reward hacks verbalized in the CoT and modify the environment to eliminate the spurious correlations. Do CoTs reliably verbalize a model’s reward hacking behavior during RL? To study this question, we construct synthetic RL environments with known reward hacks and finetune Claude 3.7 Sonnet on them. The model fully learns the reward hacks (reward > 0.99) on all RL environments, but almost never verbalizes the reward hacks ($< 2\%$ of the examples) in 5 out of 6 environments. In other words, in this experiment, CoT monitoring would not catch the majority of cases of reward hacking during RL. This result is more negative than recent results in Baker et al. (2025), which finds that CoT monitoring detects reward hacks in over 90% of generations (when the CoT is not optimized against a monitor). We think the difference in results is likely because the reward hacks studied in Baker et al. (2025) require more exploration and reasoning to exploit, and is applied to a setting (coding) where it is more common to verbalize the potential hacks.

In conclusion, our results show that:

1. CoTs of reasoning models verbalize reasoning hints at least some of the time, but rarely do so reliably (in our settings where exploiting them does not require a CoT);
2. Scaling up outcome-based RL does not steadily improve CoT faithfulness beyond a low plateau;
3. CoT monitoring may not reliably catch reward hacking during RL.

These results suggest that:

1. CoT monitoring is a valuable tool to notice unintended behaviors, and it seems most valuable when these unintended behaviors are frequent (as the proportion of generations in which the model verbalizes the unintended behavior may be low).