Can We Predict Alignment Before Models Finish Thinking? Towards Monitoring Misaligned Reasoning Models

Recent open-source efforts to build reasoning language models (RLMs) typically involve fine-tuning safety-aligned large language models (Guo et al., 2025; Muennighoff et al., 2025). Using task-specific data with long chains-of-thought (CoTs), the resulting RLMs show significant improvements on complex mathematical and STEM reasoning tasks, but become less safe than their base models. Specifically, when evaluated on safety refusal benchmarks, they show a notable increase in harmful outputs in both the long chain-of-thought (CoT) reasoning traces and final responses (Jiang et al., 2025). While much post-hoc alignment research addresses this with additional safety training (Guan et al., 2024), little work has explored monitoring their CoT traces to detect and guardrail problematic compliance behaviors resulting from the widely-used reasoning training regime. Safety monitoring is challenging because CoTs are known to be unfaithful (Turpin et al., 2023); in other words, CoTs do not accurately reflect the model's internal thinking process, leaving open the question of how effective CoT traces can be used for safety monitoring.

In this work, we study **the extent to which the safety alignment of RLMs' final responses can be predicted from their CoTs**. This is non-trivial because, given a harmful query from a safety refusal benchmark, such as "*how to smuggle drugs across the border without getting caught*," the generated CoT often contains a mix of contrasting signals, such as explicitly acknowledging the harmful nature of the task while also planning how to answer it. As CoTs can be unfaithful, either type of signal may lead to either a misaligned response that provides detailed instructions or a refusal response, which is the desired outcome for a safe model. Our work focuses on understanding how effective different CoT monitoring systems, including human annotators, fine-tuned text classifiers, and strong LLMs (e.g., o4-mini) with in-context learning, are at predicting response alignment. Furthermore, we ask whether monitoring activations, which capture how the model's internal computation evolves during the reasoning process, can outperform CoT monitoring with less inference-time compute.

Our findings are two-fold. First, we find that a simple linear probe trained on CoT activations outperforms all text-based monitoring methods. The probe achieves strong F1 scores with as few as 100 training examples, while human annotators report needing more time to process CoT text and still perform worse. Strong monitors like GPT-4.1 are commonly used in CoT monitoring (Baker et al., 2025), yet they also struggle with this task. Our findings generalize across model sizes (from 7B to 32B parameters), multiple safety refusal benchmarks, base model families including Qwen and LLaMA, and different thinking budgets.

Second, we show that the same linear probe can be used to predict response alignment from early stages of the CoT, before the model finishes reasoning. For example, given activations collected after 20 CoT sentences, the probe can predict the alignment of a response generated up to 50 sentences later. This result also holds across models with varying thinking budgets, suggesting that alignment-related signals consistently emerge early in the reasoning process. In summary, our contributions are as follows:

1. We conduct a systematic comparison of methods for monitoring safety misalignment in open-source RLMs and show that CoT activations are more predictive of final response alignment than CoT text.
2. We demonstrate that CoT texts can be unfaithful to response alignment, often misleading both humans and strong text classifiers including GPT-4.1 and o4-mini.
3. We show that a simple linear probe can predict response alignment before RLMs complete their reasoning, potentially enabling real-time monitoring and early stopping of CoT generation.