ISSUE BRIEF

FRONTIER MODEL FORUM

# Issue Brief: Chain of Thought Monitorability

**DATE**

January 27, 2026

As artificial intelligence (AI) systems become more capable, it is important that they operate safely and as intended. A common issue is the "black box" problem: we can see an AI's final output, but we often cannot understand the internal reasoning process that produced it. This opacity makes it difficult to detect, predict, and prevent potentially harmful behavior, such as deception or pursuing unintended goals.

Reasoning models and the inference-time compute paradigm change this picture. The most advanced AI models are now being trained to "think out loud" by generating a step-by-step reasoning process, known as a "Chain of Thought" (CoT), before producing a final answer. While Chain of Thought outputs are not guaranteed to perfectly reflect the model's internal computations, evidence suggests they are often meaningfully correlated with high-level reasoning processes in current frontier models. By monitoring CoT outputs, it is possible to identify harmful intent or flawed reasoning before an action is taken.

**CoT monitoring has value as a layer of defense.** There have already been numerous examples of the use of CoT monitoring to identify safety-relevant behaviors, including intentional hallucinations, reward hacking, alignment faking, and scheming. In addition, monitoring may also protect against attackers that attempt to take advantage of CoT to produce harmful outputs, either by leveraging prompt injections or other adversarial techniques. Since standard frontier AI architectures cannot do long sequences of reasoning without using the CoT, this will likely continue to hold in the near future.[1]

**Nevertheless, CoT monitoring is not a silver bullet.** CoT monitoring is most effective at protecting against harms that can only occur with

complex reasoning where the CoT itself is likely necessary.[2] By contrast, monitoring the CoT will fail to catch harms that can occur without complex reasoning. Furthermore, highly advanced models may eventually learn to deliberately hide their intentions. Therefore, CoT monitoring should be viewed as one important layer in a "defense-in-depth" approach to AI safety, complementing rather than replacing other methods.

CoT monitoring is a promising and valuable addition to the AI safety toolkit. However, it is also a fragile capability that could be easily lost through future AI development choices.[3] We recommend the following industry norms to make the most of the opportunity presented by CoT monitoring:

- **Develop and publish monitorability evaluations.** Standardized evaluations and metrics can improve assessments for how monitorable an AI model's reasoning process is. These tests would measure the clarity, coherence, and faithfulness of a model's Chain of Thought, providing a concrete score for its transparency. Once standard evaluations exist, frontier AI developers should run them on their models and report their results in model cards as appropriate.

- **All else equal, prefer not to train CoTs to "look nice."** CoT monitoring relies on us preserving the authenticity of the AI's thoughts: it would not work if the AI system's thoughts diverged from its actions. It is thus important that, where appropriate, developers avoid training the CoT for purposes other than producing better outputs and actions. As a result, we refrain from recommending the public disclosure of internal CoTs. This is to ensure developers can prioritize reasoning faithfulness without the external pressure to 'sanitize' or 'curate' thoughts for public consumption.

- **All else equal, prefer architectures that require CoTs for serial reasoning, until interpretability methods can provide similar levels of insight.** Standard frontier AI architectures, such as Transformers and Mixture of Experts, cannot perform long sequences of reasoning without using the CoT. It is plausible that CoT monitorability depends on this property, particularly in cases where we monitor hard, complex reasoning. However, there are potential architectures that would enable *latent reasoning* in an opaque, black-box manner, which would harm monitorability. While interpretability methods have the potential to replace or even exceed CoT monitoring, they do not yet meet this bar. So, the harm to CoT monitoring should be considered a significant con of latent reasoning architectures, at least until interpretability methods advance to the point of providing similar levels of insight.

- **Treat CoT monitorability as a key safety feature.** Developers should use monitorability scores to guide decisions across algorithms, data, and architectures. For example, they might choose not to use a potential

algorithmic change that would slightly increase the capability of the model, at the cost of making the CoT much more illegible (and thus less monitorable).

Importantly, several of our arguments and recommendations are simply one consideration among many, and may be overridden by others depending on the broader context. For example, process supervision has been proposed as a method that can improve safety, but may require training the CoT. It may be that future research shows that the correct approach is to use process supervision even if it degrades CoT monitorability.

In conclusion, the ability to monitor the CoT represents a significant opportunity to move beyond treating advanced AI as an impenetrable "black box", offering a powerful new tool for oversight and for ensuring that AI systems behave as intended. However, this window into the AI's reasoning is fragile, and could be closed by the technical and commercial pressures driving AI development.

Therefore, we recommend investing in research to understand CoT monitorability, developing standards to measure it, and incorporating it as a key metric in development and deployment decisions. This can ensure that we leverage this opportunity for AI safety to create an effective and ongoing safeguard.

## FOOTNOTES

1. See Figure 1 in Korbak et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety", Arxiv (2025). See also Li et al., "Chain of Thought Empowers Transformers to Solve Inherently Serial Problems", Arxiv (2024).
2. See Emmons et al., "When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors", Arxiv (2025), and Deng et al., "CoT May Be Highly Informative Despite 'Unfaithfulness'", METR (2025).
3. See Baker et al., "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation", Arxiv (2025), and Korbak et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety", Arxiv (2025).