On Learning Verifiers for Chain-of-Thought Reasoning (Extended Abstract)

Maria-Florina Balcan [1] **Avrim Blum** [2] **Zhiyuan Li** [2] **Dravyansh Sharma** [2 3]

Chain-of-Thought reasoning has emerged as a powerful approach for solving complex mathematical and logical problems. Nevertheless, it can frequently deviate from the correct path due to inaccurate or unsupported inferences. Formal mathematical reasoning, which can be validated with a formal verifier, serves as one method to mitigate this concern. However, at present, LLMs are simply not good enough to address complex problems in a formal manner, and even the task of formalizing an informal problem statement can prove to be difficult. Motivated by this fact, in this work (full version Balcan et al. 2025) we consider the problem of learning reliable verifiers for natural language Chain-of-Thought reasoning. Specifically, given a problem statement and a step-by-step solution articulated in natural language, the objective of the verifier is to produce [Yes] if all reasoning steps in the solution are valid, and [No] if they are not. In this research, we present a formal PAC-learning framework to investigate this issue. We propose and evaluate several natural verification goals, each with varying levels of strength, within this framework. Additionally, we offer upper-bounds on sample complexity for learning verifiers that meet these goals, along with lower-bound and impossibility results for acquiring other natural verification objectives without further assumptions.

As the use of LLMs to address intricate mathematical and logical challenges through chain-of-thought reasoning continues to grow, it has become essential to create verifiers capable of assessing the accuracy of these produced solutions. Specifically, despite recent progress, Chain-of-Thought (CoT) reasoning is still largely perceived to be prone to catastrophic failures due to the accumulation of errors, except in very restricted circumstances (Ling et al., 2023; Stechly et al., 2024). Detecting subtle mistakes in lengthy reasoning sequences can be particularly difficult, especially when conveyed through informal natural language expressions. This underscores the necessity for developing efficient verifiers for CoT reasoning in natural language.

In this study, we present a PAC-learning framework aimed at developing verifiers for sequential reasoners. Our learning algorithms receive a sample of various problem statements along with labeled reasoning sequences corresponding to these problems, and they must verify the accuracy of reasoning sequences that have not been previously encountered for new problems. We examine multiple related yet distinct verification objectives and study the sample complexity associated with learning verifiers that meet these requirements, providing both upper and lower bounds.

For instance, the simplest (and weakest) verification objective we examine is that, given a random reasoning trace derived from some underlying distribution D, the verifier must determine whether the reasoning is accurate or erroneous (and if erroneous, identify where the first mistake occurred), while maintaining an error rate of at most a specified $\epsilon > 0$. The goal is then to learn such a verifier from labeled data of both correct and faulty reasoning traces from the same distribution with a probability of at least $1 - \delta$. A limitation of this straightforward verification objective is its vulnerability to adaptive usage. For instance, if an LLM reasoner is informed by the verifier that a reasoning trace $x_0, x_1, ..., x_t$ is incorrect at the ith step, a typical response would be to revert and substitute x_i with an alternative step x_i' and attempt again, continuing this process until a new successful reasoning trace is discovered. However, there is no assurance that the final trace produced is accurate, due to the multiple querying rounds and the possibility that the newly queried traces may fall outside the distribution.

To tackle this challenge, we propose a more robust and reliable verification objective, wherein, given a distribution D over *problem instances*, the verifier should reject *any* faulty reasoning trace for the majority of $x_0 \sim D$. Naturally, such a verifier should also accept at least some *correct* reasoning traces from x_0, and we provide upper and lower bounds based on whether we permit the verifier to accept a designated *gold standard* reasoning trace $g(x_0)$ or if we require it to accept a significant proportion of all correct reasoning traces from x_0 without any further assumptions. These verifiers exhibit greater resilience to any distribution shifts.

Overall, we introduce a principled framework for learning verifiers for CoT reasoning. Our results complement prior work on CoT generation (Joshi et al., 2025), and achieve bounded sample complexity for learning CoT verifiers.

[1]Carnegie Mellon University [2]Toyota Technological Institute at Chicago [3]Northwestern University. Correspondence to: Dravyansh Sharma <dravy@ttic.edu>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models. Copyright 2025 by the author(s).

References

Balcan, M.-F., Blum, A., Li, Z., and Sharma, D. On learning verifiers for chain-of-thought reasoning. *arXiv preprint arXiv:2505.22650*, 2025.

Joshi, N., Vardi, G., Block, A., Goel, S., Li, Z., Misiakiewicz, T., and Srebro, N. A theory of learning with autoregressive chain of thought. *Conference on Learning Theory (COLT)*, 2025.

Ling, Z., Fang, Y., Li, X., Huang, Z., Lee, M., Memisevic, R., and Su, H. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:36407–36433, 2023.

Stechly, K., Valmeekam, K., and Kambhampati, S. Chain of thoughtlessness? An analysis of CoT in planning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.