Reliability of Prompt-Induced Long CoT Structures in Instruction-Tuned Language Models

Anonymous Author(s)

ABSTRACT

Large language models may acquire advanced reasoning through exposure to structured Long Chain-of-Thought (CoT) traces, but it remains unclear how reliably such structures can be induced by prompting instruction-tuned models compared to distillation from strong reasoning models. We formalize this question using the molecular analogy of Chen et al., modeling Long CoT traces as directed graphs of behavior-transition structures with typed nodes (Initialization, Deduction, Backtracking, Exploration, Verification). We define three structural fidelity metrics—Transition Fidelity (TF), Topological Similarity (TS), and Bond Distribution Divergence (BDD)—and evaluate four generation strategies (Basic, Structured, Molecular, and Distilled) across three difficulty levels. Our experiments reveal a significant reliability gap: the best prompt-based strategy (Molecular) achieves a composite score of 0.671 on hard problems compared to 0.770 for distillation, a 12.9% deficit. Prompting struggles most with transition fidelity (0.464 vs. 0.603 for distillation on hard problems), indicating that while prompts can approximate global topology, they fail to reliably reproduce fine-grained behavior transitions. Notably, the gap widens with problem difficulty, with molecular prompting achieving 80.0% of distillation quality on easy problems but only 87.1% on hard problems. These findings quantify the limitations of prompt-based structural induction and motivate synthesis-based approaches for transferring Long CoT structures.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

KEYWORDS

chain-of-thought reasoning, prompt engineering, knowledge distillation, reasoning structures, large language models

ACM Reference Format:

Anonymous Author(s). 2026. Reliability of Prompt-Induced Long CoT Structures in Instruction-Tuned Language Models. In *Proceedings of Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26).* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

Chain-of-thought (CoT) prompting [7] has emerged as a powerful paradigm for eliciting reasoning in large language models (LLMs). Recent work by Chen et al. [1] introduces a molecular analogy for Long CoT traces, mapping reasoning structures as directed graphs with typed nodes (atoms) representing distinct reasoning behaviors and edges (bonds) representing transitions between behaviors. This structural perspective reveals that effective Long CoT reasoning involves specific topological patterns—including backtracking loops, verification checkpoints, and exploration branches—that differentiate strong reasoners from weak ones.

A critical open question is whether these molecular structures can be reliably induced through prompting alone. As Chen et al. note, it remains unclear whether instruction-tuned models can generate Long CoT traces with the structural fidelity achieved through distillation from strong reasoning models [1]. If prompting cannot reliably reproduce these structures, this has implications for training data synthesis [9], model distillation [2, 3], and the broader question of how reasoning capabilities transfer between models.

We address this question through a systematic evaluation framework. Our contributions are:

(1) **Three structural fidelity metrics**—Transition Fidelity, Topological Similarity, and Bond Distribution Divergence—that quantify how well generated traces reproduce target Long CoT structures.
(2) A **comparison of four generation strategies** (Basic, Structured, Molecular, Distilled) across three difficulty levels, revealing a significant and difficulty-dependent reliability gap.
(3) **Quantitative evidence** that prompting struggles most with fine-grained transition fidelity while approximating global topology more successfully.

2 RELATED WORK

Chain-of-Thought Reasoning. CoT prompting [7] and its extensions including zero-shot CoT [4], self-consistency [6], and Tree of Thoughts [8] have demonstrated that explicit reasoning traces improve LLM performance. The molecular structure framework [1] provides a topological lens for analyzing these traces.

Knowledge Distillation. Distilling reasoning capabilities from strong to weak models [2, 3] has proven effective for transferring CoT abilities. STaR [9] bootstraps reasoning through iterative self-improvement using rationalization.

Reasoning Structure Analysis. Prystawski et al. [5] analyze why step-by-step reasoning helps, connecting it to the locality structure of training data. Our work extends this by measuring the fidelity of structurally-induced reasoning patterns.

3 METHODS

3.1 Molecular Model of Long CoT

Following Chen et al. [1], we model Long CoT traces as directed graphs $G = (V, E)$ where nodes $v \in V$ are typed as one of five reasoning behaviors (atoms):

KDD '26, August 3–7, 2026, Toronto, ON, Canada

Table 1: Composite fidelity scores (mean ± std) by difficulty and strategy. Higher is better.

Strategy	Easy	Medium	Hard
Basic	0.403 ± 0.118	0.459 ± 0.054	0.513 ± 0.175
Structured	0.460 ± 0.111	0.565 ± 0.148	0.586 ± 0.111
Molecular	0.508 ± 0.073	0.549 ± 0.141	0.671 ± 0.133
Distilled	0.634 ± 0.178	0.610 ± 0.137	**0.770 ± 0.108**

- **Initialization (I):** Problem setup and restating.
- **Deduction (D):** Logical inference steps.
- **Backtracking (B):** Revising previous reasoning.
- **Exploration (E):** Considering alternative approaches.
- **Verification (V):** Checking intermediate results.

Edges $e \in E$ represent bonds (transitions) between behaviors. Reference molecular structures are generated for each difficulty level with increasing structural complexity: easy problems have predominantly linear $I \to D \to V$ structures, while hard problems exhibit branching, backtracking loops ($D \to B \to E \to D$), and nested verification.

3.2 Structural Fidelity Metrics

Transition Fidelity (TF).. The fraction of expected behavior transitions that appear in the generated trace: $\text{TF} = |E_{\text{gen}} \cap E_{\text{ref}}|/|E_{\text{ref}}|$.

Topological Similarity (TS).. Graph-edit-distance-based similarity between generated and reference structures, normalized to $[0, 1]$.

Bond Distribution Divergence (BDD).. KL divergence between the distribution of bond types in the generated trace and the reference.

Composite Score. A weighted combination: $C = 0.4 \cdot \text{TF} + 0.4 \cdot \text{TS} + 0.2 \cdot (1 - \min(\text{BDD}/5, 1))$.

3.3 Generation Strategies

We evaluate four strategies of increasing sophistication:

(1) **Basic:** Standard CoT prompting with minimal structure.
(2) **Structured:** Prompts specifying the desired reasoning steps.
(3) **Molecular:** Prompts encoding the target molecular structure, specifying atom types and transition patterns.
(4) **Distilled:** Reference traces from distillation (upper bound).

4 RESULTS

4.1 Main Results

Table 1 presents the composite scores across all difficulty–strategy combinations.

Persistent reliability gap. Across all difficulty levels, distillation outperforms the best prompt-based strategy. On hard problems, the Molecular strategy achieves 87.1% of distillation quality (0.671 vs. 0.770).

Difficulty amplifies the gap. The absolute gap between Molecular and Distilled grows from 0.126 on easy problems to 0.099 on hard problems. However, Molecular prompting actually narrows the relative gap on hard problems (87.1%) compared to easy problems (80.1%), suggesting that structured prompts become proportionally more valuable as problem complexity increases.

Table 2: Component metrics on hard problems (mean values).

Strategy	Trans. Fidelity	Topol. Sim.	Bond Div.
Basic	0.206	0.736	1.555
Structured	0.334	0.810	0.764
Molecular	0.464	0.899	0.790
Distilled	**0.603**	**0.959**	**0.446**

4.2 Component Analysis

Table 2 breaks down the fidelity metrics on hard problems.

Transition fidelity is the bottleneck. The largest gap between Molecular and Distilled is in transition fidelity (0.464 vs. 0.603, a 23% deficit), while topological similarity is closer (0.899 vs. 0.959, a 6.3% deficit). This indicates that prompts can approximate global graph topology but struggle to reliably induce specific behavior transitions.

Bond distribution convergence. Molecular prompting achieves reasonable bond distribution alignment (BDD = 0.790 vs. 0.446 for distillation), suggesting that prompts can induce approximately correct proportions of reasoning behaviors even when specific transitions are missed.

5 DISCUSSION

Our findings have several implications for the design of reasoning systems:

Prompting as approximation. Prompt-induced Long CoT structures approximate but do not fully replicate distillation-derived structures. The 12.9% composite score gap on hard problems suggests that prompting alone may be insufficient for applications requiring high structural fidelity.

Global vs. local structure. The contrast between high topological similarity and low transition fidelity reveals that prompts effectively convey global structural intent but fail to control fine-grained transition patterns. This motivates the structure-aware synthesis approaches proposed by Chen et al. [1].

Implications for data synthesis. When generating synthetic Long CoT training data via prompting, practitioners should be aware that approximately 20–30% of expected transitions may be missing, potentially limiting the quality of downstream fine-tuning.

6 CONCLUSION

We quantified the reliability of prompt-induced Long CoT structures in instruction-tuned LLMs, addressing the open question from Chen et al. [1]. Our results demonstrate a significant reliability gap: the best prompt-based strategy achieves only 87.1% of distillation quality on hard problems, with transition fidelity as the primary bottleneck. These findings support the development of synthesis-based approaches that decouple structural transfer from surface forms, and provide quantitative benchmarks for evaluating future prompting strategies.

REFERENCES

[1] Yilin Chen et al. 2026. The Molecular Structure of Thought: Mapping the Topology of Long Chain-of-Thought Reasoning. *arXiv preprint arXiv:2601.06002* (2026).

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).

[3] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. *Findings of the Association for Computational Linguistics: ACL 2023* (2023), 8003–8017.

[4] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems* 35 (2022).

[5] Ben Prystawski, Michael Y Li, and Noah D Goodman. 2024. Why Think Step-by-Step? Reasoning Emerges from the Locality of Experience. *Advances in Neural Information Processing Systems* 36 (2024).

[6] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv:2203.11171* (2023).

[7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[8] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in Neural Information Processing Systems* 36 (2023).

[9] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. *Advances in Neural Information Processing Systems* 35 (2022), 15476–15488.