Theorem-of-Thought: A Multi-Agent Framework for Abductive, Deductive, and Inductive Reasoning in Language Models

Samir Abdaljalil[1], Hasan Kurban[2*], Khalid Qaraqe[2], Erchin Serpedin[1]
[1]Electrical & Computer Engineering, Texas A&M University, College Station, TX, USA
[2]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

Abstract

Large language models (LLMs) have shown strong performance across natural language reasoning tasks, yet their reasoning processes remain brittle and difficult to interpret. Prompting techniques like Chain-of-Thought (CoT) enhance reliability by eliciting intermediate reasoning steps or aggregating multiple outputs. However, they lack mechanisms for enforcing logical structure and assessing internal coherence. We introduce **Theorem-of-Thought (ToTh)**, a novel framework that models reasoning as collaboration among three parallel agents, each simulating a distinct mode of inference: abductive, deductive, and inductive. Each agent produces a reasoning trace, which is structured into a formal reasoning graph. To evaluate consistency, we apply Bayesian belief propagation guided by natural language inference (NLI), assigning confidence scores to each step. The most coherent graph is selected to derive the final answer. Experiments on symbolic (WebOfLies) and numerical (MultiArith) reasoning benchmarks show that ToTh consistently outperforms CoT, Self-Consistency, and CoT-Decoding across multiple LLMs, while producing interpretable and logically grounded reasoning chains. Our findings suggest a promising direction for building more robust and cognitively inspired LLM reasoning. The implementation is available at https://github.com/KurbanIntelligenceLab/theorem-of-thought.

1 Introduction

Large language models (LLMs) have achieved impressive performance across a wide range of natural language understanding and generation tasks (Wang et al., 2024), enabled by advances in in-context learning (Sia et al., 2024), instruction tuning (Zhang et al., 2024), and chain-of-thought (CoT) prompting (Wei et al., 2022). These methods have extended LLMs' capabilities to handle complex forms of reasoning, including mathematical, logical, and commonsense inference.

Despite these advances, LLM reasoning remains shallow and unreliable. Existing approaches often rely on single-shot or sampling-based decoding along linear reasoning paths, making them susceptible to hallucinations (Abdaljalil et al., 2025), logical inconsistencies (Uceda Sosa et al., 2024), and weak generalization (Liu et al., 2025). Methods such as CoT and Self-Consistency (Wei et al., 2022; Wang et al., 2023) encourage intermediate steps and majority voting across sampled outputs, but lack mechanisms to verify internal coherence and model the logical structure of reasoning. As a result, outputs may appear fluent and plausible while remaining logically unsound.

This brittleness contrasts sharply with human reasoning, which is inherently multifaceted. Drawing on insights from cognitive science (Okoli, 2022), we observe that human inference typically blends three complementary modes—abduction, deduction, and induction—that support explanation, derivation, and generalization. However, LLMs typically conflate these distinct processes into a single, undifferentiated flow, limiting both interpretability and reliability.

To address this gap, we propose **Theorem-of-Thought (ToTh)**, a framework that models diverse reasoning strategies through structured, verifiable interactions. ToTh employs three specialized agents, each emulating a distinct cognitive mode:

- **Abduction**: inferring plausible explanations for observed facts;
- **Deduction**: deriving valid conclusions from given premises;
- **Induction**: generalizing from patterns or examples.

*Corresponding author: hkurban@hbku.edu.qa

Each agent independently generates a reasoning trace, which is transformed into a Formal Reasoning Graph (FRG)—a directed graph where nodes represent intermediate conclusions and edges capture logical dependencies. We evaluate the internal consistency of each FRG using Bayesian belief propagation, with edge confidence scores calibrated via a Natural Language Inference (NLI) model. A composite score balancing average belief and logical entropy is used to select the most coherent graph, from which the final answer is extracted.

Contributions. The key results of this work are:

- We introduce ToTh, a structured reasoning framework that integrates abductive, deductive, and inductive inference into a modular LLM-based pipeline.
- We develop a belief propagation mechanism over reasoning graphs, leveraging NLI to assess and score logical coherence through Bayesian updates.
- We demonstrate that ToTh consistently outperforms state-of-the-art reasoning methods (e.g., CoT, Self-Consistency, CoT-Decoding) across multiple LLMs.
- Our evaluation on symbolic (WEBOFLIES) and numerical (MULTIARITH) benchmarks highlights ToTh's robustness on tasks requiring multi-step inference—settings where direct prompting often fails (Allen-Zhu and Li, 2025).

The remainder of the paper is organized as follows: Section 2 reviews related work. Section 3 presents the ToTh framework. Section 4 describes the experimental setup, and Section 5 analyzes the results obtained. Section 6 concludes with implications for structured reasoning in LLMs and future research directions.

2 Related Work

Prompt-based Reasoning in LLMs. A growing body of work explores prompting strategies to enhance the reasoning capabilities of LLMs. CoT prompting (Wei et al., 2022) encourages models to decompose problems into intermediate steps, guiding reasoning along a linear path. Building on this, Auto-CoT (Zhang et al., 2023) automates prompt generation by sampling diverse questions and producing corresponding reasoning traces, reducing manual effort. Beyond prompt generation, several works focus on optimizing prompt selection strategies. ActivePrompt (Diao et al., 2024) identifies high-uncertainty instances for annotation, improving data efficiency and reasoning robustness through active learning. More recent approaches introduce explicit structure into the reasoning process. Tree-of-Thought (ToT) (Yao et al., 2023) enables multi-path exploration with internal evaluation, while Graph-of-Thought (GoT) (Yao et al., 2024) structures reasoning as a graph to better model dependencies between steps.

Instruction Tuning for Reasoning. Instruction tuning and knowledge distillation offer alternative approaches to eliciting reasoning in LLMs without relying on explicit prompts (Lobo et al., 2025; Ranaldi and Freitas, 2024; Lai and Nissim, 2024). While effective, these methods typically require computationally intensive fine-tuning on large-scale datasets annotated with reasoning traces and CoT examples, which are often costly and domain-specific. Recent work has explored more indirect supervision strategies. For instance, Liu et al. (2024) introduce proxy tuning, which leverages auxiliary models to contrast a base LLM with its adapted variant. Although this approach reduces the need for direct supervision, it still assumes access to CoT-like outputs and pre-aligned reasoning benchmarks.

3 Methodology

ToTh is a graph-based reasoning framework designed to enhance the accuracy, interpretability, and generalization capabilities of LLMs on complex tasks. It decomposes reasoning into three modular agents, each simulating a classical inference paradigm—abduction, deduction, and induction. Each agent produces a structured reasoning trace, which is composed into a FRG. Final answers are derived via NLI-calibrated Bayesian belief propagation and composite graph scoring. The full pipeline is depicted in Fig. 1.

ToTh differs from prior reasoning paradigms along three axes: architecture, supervision, and verification. Prompt-based methods (e.g., CoT, ToT, GoT) elicit reasoning via linear or loosely structured traces, yet lack mechanisms for enforcing logical consistency. Instruction-tuned models embed reasoning behavior through fine-tuning on annotated traces, often requiring large datasets and

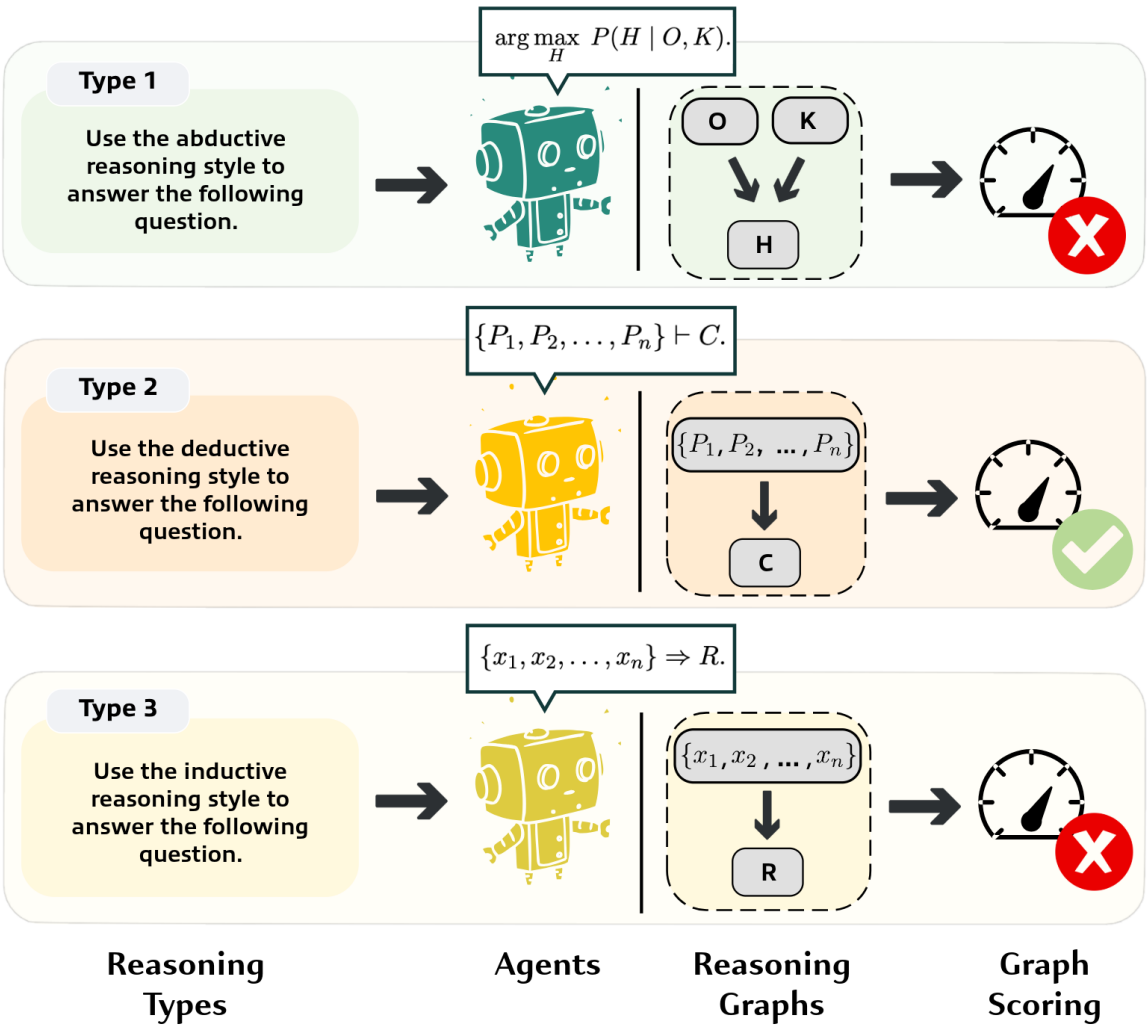

Figure 1: Overview of the Theorem-of-Thought (ToTh) reasoning pipeline. A question is independently processed by three agents, each using a distinct reasoning style: abductive (Type 1), deductive (Type 2), and inductive (Type 3). Each agent produces a structured reasoning graph, which is scored via Bayesian confidence propagation. Abduction infers the best hypothesis H given observations O and knowledge K (i.e., $\arg\max_H P(H \mid O, K)$); deduction derives a conclusion C from premises $\{P_1, \ldots, P_n\}$ (i.e., $\{P_i\} \vdash C$); induction generalizes from examples $\{x_1, \ldots, x_n\}$ to a rule R (i.e., $\{x_i\} \Rightarrow R$). The highest-scoring graph contributes its final node as the answer. ✓ and ✗ indicate whether a given agent's output was selected.

remaining opaque at inference time. While both families reflect growing interest in structured multi-step reasoning, they typically operate within monolithic or implicit architectures and do not support formal consistency checking. In contrast, ToTh instantiates distinct cognitive agents, integrates their outputs into an interpretable graph, and explicitly verifies reasoning coherence through NLI-guided Bayesian inference—enabling modular, transparent, and verifiable reasoning beyond the scope of existing methods.

Multi-Paradigm Reasoning Agents. Given a natural language question q, ToTh deploys three independent solver agents, each aligned with a distinct classical mode of inference: abductive, deductive, and inductive reasoning. These paradigms are formally defined as follows.

The abductive reasoning agent a_1 infers the most plausible hypothesis H given a set of observations O and background knowledge K, formalized as:

$$a_1 : \quad \arg\max_H \; P(H \mid O, K).$$

The deductive reasoning agent a_2 derives a conclusion C that logically follows from a set of premises $\{P_1, P_2, \ldots, P_n\}$, represented as:

$$a_2 : \quad \{P_1, P_2, \ldots, P_n\} \vdash C.$$

The inductive reasoning agent a_3 generalizes a rule

R from observed examples $\{x_1, x_2, \ldots, x_n\}$, expressed as:

$$a_3: \quad \{x_1, x_2, \ldots, x_n\} \Rightarrow R.$$

Each agent $a_i \in \{a_1, a_2, a_3\}$ independently produces a reasoning trace

$$\mathbf{r}^{(i)} = \left[r_1^{(i)}, r_2^{(i)}, \ldots, r_{s_i}^{(i)}\right],$$

where $r_j^{(i)}$ denotes the j-th step in the agent's reasoning process.

Formal Reasoning Graph Construction. Each reasoning trace $\mathbf{r}^{(i)}$ is transformed into a directed graph $G^{(i)} = (V^{(i)}, E^{(i)})$, where $V^{(i)}$ denotes the set of nodes representing individual reasoning steps, and $E^{(i)}$ represents directed edges encoding inferential relationships between those steps. Edges $(v_u \to v_v) \in E^{(i)}$ are inferred using a pre-trained NLI model, which assesses the semantic relationship between reasoning steps. Each edge is annotated with a trust score $\theta_{uv} \in [0, 1]$ based on the predicted label:

$$\theta_{uv} = \begin{cases} 0.95 & \text{if entailment} \\ 0.60 & \text{if neutral} \\ 0.10 & \text{if contradiction} \end{cases}$$

These scores quantify the strength of logical entailment between intermediate steps, providing a calibrated basis for probabilistic reasoning in the subsequent belief propagation stage.

Bayesian Confidence Propagation. To model belief flow across the graph, belief values are propagated using a Bayesian update rule, adapted from classical formulations of belief propagation in probabilistic graphical models (Pearl, 1988).

Each node $v \in V$ is initialized with a prior confidence $P(v) = 0.5$, reflecting maximum uncertainty. For a node v_c with a single parent v_p and associated trust score θ_{pc}, the updated belief is computed using a Bayesian update rule:

$$P(v_c) = \frac{P(v_p) \cdot \theta_{pc}}{P(v_p) \cdot \theta_{pc} + (1 - P(v_p)) \cdot (1 - \theta_{pc})}.$$

In the case of multiple parents $\{v_{p_1}, \ldots, v_{p_m}\}$, the belief for v_c is computed as the average of individual updates from each parent:

$$P(v_c) = \frac{1}{m} \sum_{j=1}^{m} f\big(P(v_{p_j}), \theta_{p_j c}\big)$$

$$f(p, \theta) = \frac{p \cdot \theta}{p \cdot \theta + (1 - p)(1 - \theta)}.$$

This recursive formulation propagates confidence through the graph, amplifying agreement across consistent reasoning paths while attenuating belief when upstream uncertainty or contradiction is detected.

Graph Scoring. Each reasoning graph $G^{(i)}$ is evaluated based on a trade-off between average node confidence and logical uncertainty. We prioritize graphs that are both confident (high belief) and low in uncertainty (low entropy). The mean confidence is computed as

$$\mu^{(i)} = \frac{1}{|V^{(i)}|} \sum_{v \in V^{(i)}} P(v),$$

and the normalized binary entropy is given by

$$H^{(i)} = -\frac{1}{|V^{(i)}|} \sum_{v \in V^{(i)}} h(P(v))$$

$$h(p) = p \log p + (1 - p) \log(1 - p).$$

The final score combines both terms:

$$\text{Score}(G^{(i)}) = \mu^{(i)} - H^{(i)}.$$

The reasoning graph with the highest score is selected as the final candidate:

$$G^* = \arg\max_i \text{Score}(G^{(i)}).$$

Answer Extraction. The final answer is extracted from the terminal node of the selected graph G^*, corresponding to the last step in the associated reasoning trace.

Theoretical Complexity. Let $k = 3$ denote the number of reasoning agents, and s the number of reasoning steps generated per agent. The ToTh framework involves three main stages of computation: trust estimation, belief propagation, and graph scoring. During trust estimation, each agent produces a sequence of reasoning steps, and an NLI model is applied to each adjacent pair to evaluate the strength of logical connection. Since each trace contains at most $s - 1$ such pairs, the total number of NLI evaluations across all agents is $\mathcal{O}(k \cdot s)$. In the belief propagation stage, each node in the constructed reasoning graphs is visited exactly once in topological order, and its posterior confidence is updated based on incoming trust scores using

a Bayesian update rule, resulting in $\mathcal{O}(k \cdot s)$ total updates. Finally, graph scoring involves computing the average confidence and entropy over all nodes in each graph, which also requires $\mathcal{O}(k \cdot s)$ time. Therefore, the end-to-end complexity of the ToTh pipeline is $\mathcal{O}(k \cdot s)$, linear in both the number of agents and the number of reasoning steps per agent.

This makes ToTh substantially more efficient than sampling-based methods such as Self-Consistency or CoT-Decoding, which require $\mathcal{O}(n)$ decoding passes, where n is the number of sampled reasoning chains. In contrast, ToTh executes a single, structured reasoning pass per agent, followed by lightweight verification and scoring, offering a more scalable and interpretable alternative to stochastic decoding.

4 Experiments

Data. ToTh was evaluated on two representative reasoning benchmarks. MULTIARITH (Roy et al., 2015) targets compositional numerical inference through multi-step arithmetic word problems. WEBOFLIES (Suzgun et al., 2023), part of the BIG-BENCH-HARD suite, involves determining truth values among logically entangled symbolic statements. These datasets are known to challenge LLMs under direct prompting (Allen-Zhu and Li, 2025), making them suitable for testing structured reasoning capabilities.

Models. Three publicly available LLMs were selected to provide diversity in scale, alignment, and architecture: (1) MISTRAL-7B (Jiang et al., 2023)[1], a general-purpose decoder model with efficient scaling; (2) DEEPSEEK-7B (DeepSeek-AI et al., 2025)[2], an instruction-tuned model optimized for multi-turn reasoning and alignment; and (3) PHI-3.5 MINI (Abdin et al., 2024)[3], a lightweight model designed for educational, low-cost reasoning tasks. This selection spans compact inference-efficient models to instruction-aligned reasoning-focused systems.

Baselines. ToTh was compared with three strong baselines: CoT (Wei et al., 2022), Self-Consistency (Wang et al., 2023), and CoT-Decoding (Wang and Zhou, 2024). CoT prompts the model to generate intermediate reasoning steps before answering. Self-Consistency improves robustness by sampling $n = 20$ completions and selecting the most frequent answer. CoT-Decoding eliminates explicit prompting by using diverse decoding paths to stimulate latent reasoning behaviors.

Experimental Setup. All models were evaluated in their released form without fine-tuning. Decoding was performed with temperature 0.7 and a maximum output length of 526 tokens. RoBERTa-MNLI[4] was used for scoring reasoning coherence, consistent with prior work on NLI-based output validation (Farquhar et al., 2024). Inputs were uniformly formatted as "Q: [question] \n A:" across all methods for consistency with baselines (Wang and Zhou, 2024).

To direct reasoning behavior, the following instruction was prepended to each input, with the appropriate `{style}` keyword for each agent:

> `Use the` **`{style}`** `reasoning style to answer the following question. Follow these instructions carefully:`
>
> - `Break the problem into clear, numbered reasoning steps using {style}.`
> - `Reference any known principles, patterns, or assumptions involved.`
> - `Arrive at a final answer that directly responds to the question.`

All experiments used a single decoding pass per input. Random seeds were fixed, and decoding settings were held constant for reproducibility.

5 Results

5.1 Main Experimental Results

Results are reported as answer accuracy (%) and summarized in Figure 2.

Performance Across Models. ToTh consistently outperforms all baseline methods on both tasks when evaluated with MISTRAL-7B and DEEPSEEK-7B, demonstrating clear gains in reasoning accuracy. On PHI-3.5 MINI, although CoT-Decoding marginally surpasses ToTh on certain

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[2] https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat

[3] https://huggingface.co/microsoft/Phi-3.5-mini-instruct

[4] https://huggingface.co/FacebookAI/roberta-large-mnli

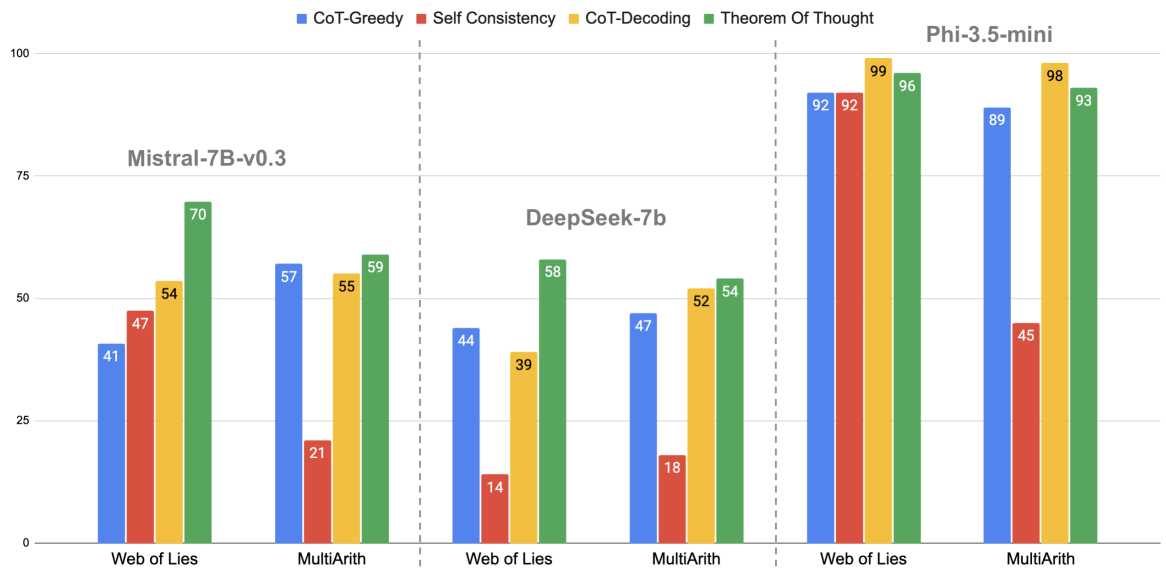

Figure 2: Accuracy (%) comparison across reasoning pipelines on two benchmark tasks (WEBOFLIES and MULTIARITH) using three open-source language models: MISTRAL-7B-V0.3, DEEPSEEK-7B, and Phi-3.5-mini. Each group of bars corresponds to a different reasoning method: CoT-Greedy (blue), Self-Consistency (red), CoT-Decoding (yellow), and our proposed Theorem-of-Thought (green).

instances, ToTh maintains consistently strong performance across both symbolic and numerical tasks. For example, on the WEBOFLIES dataset, ToTh improves over CoT-Greedy by 29% and 14% on MISTRAL-7B and DEEPSEEK-7B, respectively, and remains within 3% of the top-performing method on PHI-3.5 MINI. These results highlight ToTh's robustness and generalization across models of varying scale and alignment.

Comparison with CoT-Decoding. While CoT-Decoding performs strongly on Phi-3.5-mini, achieving near-perfect scores on WEBOFLIES (99%), ToTh achieves comparable or slightly lower performance (96%) while maintaining higher consistency across models. For example, on the MULTIARITH dataset, ToTh surpasses CoT-Decoding by 4–5 points on both MISTRAL-7B and DEEPSEEK-7B, indicating superior generalization in numerical reasoning.

Self-Consistency Under-performance. Surprisingly, Self-Consistency under-performs across all settings, particularly on symbolic tasks. For instance, it yields only 14% and 21% on WEBOFLIES and MULTIARITH with DEEPSEEK-7B and MISTRAL-7B, respectively. This suggests that majority-vote over stochastic generations fails to capture structured dependencies, especially in logic-heavy tasks.

Model Sensitivity. As expected, performance scales with model capability. Phi-3.5-mini achieves the highest absolute scores across all methods, reflecting its stronger alignment and training. However, ToTh's margin over baselines remains meaningful even at lower model scales, suggesting that the architecture contributes to reasoning robustness beyond just model size. While DEEPSEEK-7B is trained with reasoning capabilities in mind, its broader training objectives, including code generation and open-ended question answering, may diffuse its specialization in structured reasoning tasks. In contrast, Phi-3.5-mini benefits from a targeted curriculum focused on educational and step-by-step problem-solving, which likely accounts for its superior performance on both symbolic and mathematical benchmarks. Interestingly, MISTRAL-7B consistently outperforms DEEPSEEK-7B despite being similar in size. This may be attributed to Mistral's cleaner, reasoning-focused pretraining data and architecture-level optimizations, which enhance its ability to follow multi-step instructions and maintain logical coherence across token spans.

5.2 Robustness Under Reasoning Complexity

To evaluate the robustness of ToTh under increasing reasoning complexity, experiments were conducted using the MISTRAL-7B model on both symbolic and numerical tasks. Table 1 presents

	WEBOFLIES			MULTIARITH		
	3	4	5	d_0/l_3	d_0/l_4	d_2/l_3
CoT-G	41	32	19	57	26	14
SelfC	48	47	38	21	6	17
CoT-Dec	54	48	**46**	55	41	**24**
ToTh	**70**	**56**	43	**59**	**45**	21

Table 1: Accuracy (%) of MISTRAL-7B on symbolic (WEBOFLIES) and mathematical (MULTIARITH) reasoning tasks across increasing levels of difficulty. Columns 3–5 correspond to symbolic reasoning with 3, 4, and 5 interdependent statements, respectively. Columns d_0/l_3, d_0/l_4, and d_2/l_3 represent arithmetic reasoning problems categorized by depth and length: d denotes operation depth and l indicates sequence length. ToTh achieves the highest accuracy in 5 out of 6 settings and remains competitive even on the most complex instances, demonstrating consistent performance across symbolic and numerical domains. **Bold**: best performance; Underlined: second-best.

accuracy results stratified by problem difficulty: the number of interdependent statements (3–5) for WEBOFLIES, and operation depth/length combinations for MULTIARITH.

ToTh maintains strong performance across all difficulty levels, outperforming or closely matching leading baselines. In symbolic reasoning, ToTh achieves 43% accuracy on the most challenging setting (5 statements), significantly exceeding CoT-Greedy (19%) and Self-Consistency (38%), and closely approaching CoT-Decoding (46%). This trend persists across simpler instances, where ToTh attains the highest scores at 3 and 4 statements.

For numerical reasoning, ToTh delivers the strongest results at lower complexity levels—achieving state-of-the-art performance at d_0/l_3 (59%) and d_0/l_4 (45%)—and remains competitive even at higher complexity (d_2/l_3), with accuracy comparable to CoT-Decoding (21% vs. 24%). These findings highlight ToTh's capacity to generalize across task difficulty and suggest that its structured, multi-agent reasoning design offers a scalable advantage under increased inference load.

6 Conclusion and Future Work

This work presents Theorem-of-Thought (ToTh), a graph-based reasoning framework that integrates abductive, deductive, and inductive inference through a modular multi-agent design. Each agent generates structured reasoning traces, which are composed into formal graphs and verified using NLI-calibrated Bayesian confidence propagation. This approach supports both accurate prediction and interpretable, logically grounded reasoning. Empirical evaluations on symbolic and numerical benchmarks demonstrate that ToTh consistently outperforms strong prompting and decoding baselines, particularly in scenarios requiring structured logical inference.

ToTh introduces a new paradigm in reasoning with language models by treating inference as a verifiable, compositional process, rather than a monolithic generation task. Future research will explore dynamic agent routing based on input characteristics, inter-agent collaboration protocols, and adaptive trust estimation via fine-tuned and ensemble-based NLI models. Extending the framework to scientific hypothesis validation, law and policy reasoning, and multimodal domains such as visual question answering represents a promising direction for advancing general-purpose, verifiable reasoning in large language models.

Limitations

Fixed Reasoning Types. ToTh presumes a uniform decomposition into abductive, deductive, and inductive reasoning across all inputs. While this modularity improves interpretability, it imposes a fixed cognitive scaffold that may not align with tasks requiring hybrid or atypical inference patterns. For example, creative tasks or ambiguous prompts may benefit from dynamically blending reasoning types or emphasizing one over others. This rigidity can limit ToTh's adaptability and lead to suboptimal trace composition in such cases. Future work may explore data-driven and context-sensitive agent routing, allowing the framework to selectively instantiate and suppress reasoning paradigms based on input semantics.

Propagation Sensitivity. The Bayesian confidence propagation mechanism is sensitive to noise in low-confidence nodes, which may attenuate otherwise valid reasoning chains or distort belief estimates in deeper regions of the graph. This can occur in longer traces where errors in early reasoning steps propagate disproportionately, reducing the reliability of final predictions. Moreover, current propagation is uniform and unregularized, lacking robustness mechanisms against adversarial and inconsistent intermediate steps. Incorporating calibrated uncertainty modeling, edge dropout, and confidence smoothing—potentially informed by fine-grained entailment distributions—could enhance stability and mitigate the amplification of localized inconsistencies.

References

Samir Abdaljalil, Hasan Kurban, Parichit Sharma, Erchin Serpedin, and Rachad Atat. 2025. Sindex: Semantic inconsistency index for hallucination detection in llms. *Preprint*, arXiv:2503.05980.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Zeyuan Allen-Zhu and Yuanzhi Li. 2025. Physics of language models: Part 3.2, knowledge manipulation. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR 2025 Poster.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2024. Active prompting with chain-of-thought for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1350, Bangkok, Thailand. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630. © 2024. The Author(s).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026, Bangkok, Thailand. Association for Computational Linguistics.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. 2024. Tuning language models by proxy. In *Proceedings of the Conference on Language Modeling (COLM)*.

Chaoqun Liu, Qin Chao, Wenxuan Zhang, Xiaobao Wu, Boyang Li, Anh Tuan Luu, and Lidong Bing. 2025. Zero-to-strong generalization: Eliciting strong capabilities of large language models iteratively without gold labels. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3716–3731, Abu Dhabi, UAE. Association for Computational Linguistics.

Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. 2025. On the impact of fine-tuning on chain-of-thought reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11679–11698, Albuquerque, New Mexico. Association for Computational Linguistics.

Chitu Okoli. 2022. Inductive, abductive and deductive theorizing. *SSRN Electronic Journal*, Forthcoming.

Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Leonardo Ranaldi and Andre Freitas. 2024. Self-refine instruction-tuning for aligning reasoning in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2347, Miami, Florida, USA. Association for Computational Linguistics.

Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

Suzanna Sia, David Mueller, and Kevin Duh. 2024. Where does in-context learning happen in large language models? In *Proceedings of the 2024 Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS 2024 Poster.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Rosario Uceda Sosa, Karthikeyan Natesan Ramamurthy, Maria Chang, and Moninder Singh. 2024. Reasoning about concepts with llms: Inconsistencies abound. *Conference on Language Models (COLM)*. Published: 10 Jul 2024, Last Modified: 26 Aug 2024.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *Proceedings of the 2024 Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS 2024 Poster.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yao Yao, Zuchao Li, and Hai Zhao. 2024. GoT: Effective graph-of-thought reasoning in language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2901–2921, Mexico City, Mexico. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *Proceedings of International Conference on Learning Representations (ICLR)*. Poster.