When Reasoning Collapses: A Depth-Aware Probe into LLM Reasoning (Student Abstract)

Azka Ikramullah[1], Abdul Majeed[2], Kyunghyun Lee[2], Seong Oun Hwang[2*]

[1] Department of IT Convergence Engineering, Gachon University, South Korea
[2] Department of Computer Engineering, Gachon University, South Korea
202540395@gachon.ac.kr, ab09@gachon.ac.kr, kyunghyunlee@gachon.ac.kr, sohwang@gachon.ac.kr

Abstract

Large language models (LLMs) often perform better when prompted to explain their reasoning, but it remains unclear how well such gains persist as reasoning depth increases. In this work, we propose a depth-aware evaluation framework alongside the performance results on two structured datasets: CLUTRR (kinship reasoning) and ProofWriter (logical entailment), comparing direct vs. reasoning (reasoning depth = number of inference steps required) prompts across five models. Reasoning gave small gains at shallow depths but quickly weakened and often reversed as tasks grew more complex. In ProofWriter, GPT-5 reached 90% accuracy at depth four in direct model, yet its reasoning accuracy fell below baseline after depth two. Smaller open-source models showed only unstable or negligible gains, underscoring that reasoning in LLMs remains brittle with increased depth.

Introduction

Large language models (LLMs) are increasingly used in domains that require reasoning, yet their reliability in depth remains unclear (reasoning depth = number of inference steps required). The chain-of-thought prompting improves performance on shallow benchmarks (Wei et al. 2022), studies show LLMs often give only an illusion of reasoning and break down on complex tasks (Shojaee* et al. 2025). Without such evaluations, model output may collapse as complexity increases. To address this gap, we introduce a depth-aware framework designed to test whether the benefits of reasoning prompts persist as depth increases. This work moves beyond shallow benchmarks by explicitly exposing depth-driven brittleness in LLM reasoning.

Methodology and Results

Motivated by the concerns of (Shojaee et al. 2025), our framework contrasts direct vs. reasoning prompts across controlled inference depths. Figure 1 illustrates the evaluation framework. We use two datasets: CLUTRR (Sinha et al. 2019) for kinship reasoning (depths 2–6) and ProofWriter (Tafjord and Clark 2021) for logical entailment (depths 1–5). Each item is instantiated in two forms: a *direct* prompt requesting only the final label, and a *reasoning* prompt allowing intermediate steps before the answer. The framework has three modules: (i) balanced sampling per depth, (ii) prompt generation, and (iii) model evaluation with accuracy and collapse-depth metrics.

This setup is lightweight because it relies on only 100 balanced items per dataset, making replication feasible while still spanning a meaningful range of depths. We selected 100 items to strike a balance between breadth and interpretability, avoiding evaluation fatigue while ensuring consistent results. Collapse depth is defined at 50% accuracy, marking the point where performance drops below a minimally useful threshold.

We evaluate five models—GPT-5 (closed-source), DeepSeek-32B, GLM-4.5-Air, LLaMA-3.1-8B, and Qwen-2.5-7B (open-source). Model outputs are reduced to a single predicted label and scored against gold answers. Accuracy is reported by depth, and collapse depth indicates the first depth where accuracy falls below 50%. This setup enables consistent comparisons between direct and reasoning prompts across both open-source and closed-source settings.

Key Findings

Our experiments highlight three findings:

1. **Brief gains followed by collapse.** On CLUTRR, GPT-5 improves by +10 pp at depth two but drops off beyond depth three. On ProofWriter, it reaches 90% accuracy at depth four with direct answers, yet reasoning accuracy falls from 20% at depth one to 0% by depth four.

2. **Modest or unstable benefits in open models.** LLaMA-3.1-8B shows small but consistent gains, rising from 20% to 30% at depth two. Qwen-2.5-7B produces sharper but less stable spikes, such as a +20 pp gain at CLUTRR depth four, followed by collapse.

3. **Models that do not benefit from reasoning.** DeepSeek-32B and GLM-4.5-Air are ineffective across depths, showing no improvements from reasoning prompts.

Trends and Failures

Figure 2 illustrates how accuracy shifts with depth. GPT-5 performs well with direct answers on ProofWriter, but

*Corresponding Author

Model	Best Dir/Rea (PW)	Best Gain (CL)	Coll. (CL/PW)	d2	d3	d4	d5
GPT-5	90/0%@d4	+10%@d2	2/1	60/20/+10	80/10/-10	90/0/-10	80/0/-10
LLaMA-3.1-8B	60/50%@d5	+10%@d2	2/2	20/20/+10	50/50/0	40/50/0	60/50/0
Qwen-2.5-7B	60/50%@d5	+20%@d4	2/1	30/20/0	40/40/0	30/40/20	60/50/-30
DeepSeek-32B	0%	$\leq$10%	2/1	0/0	0/0	0/0	0/0
GLM-4.5-Air	0%	0%	2/1	0/0	0/0	0/0	0/0

Table 1: Comparative analysis of five LLMs: reports best direct accuracy (PW, ProofWriter)/reasoning accuracy, best gain (CL, CLUTRR), collapse depths (Coll.), and per-depth accuracy (% direct/reasoning/gain) across models.

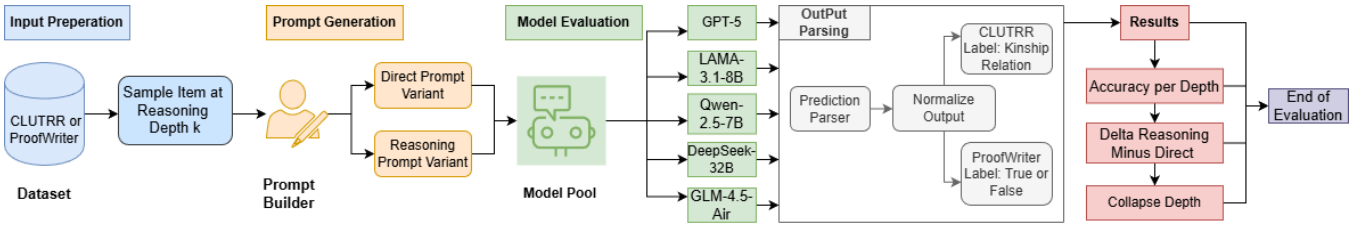

Figure 1: Evaluation framework. At each depth k, we pair direct and reasoning prompts, evaluate five models, and reduce outputs to a single label. Tracking accuracy, reasoning gains, and collapse depth reveal how reasoning robustness deteriorates as complexity grows, offering a simple yet principled probe beyond shallow benchmarks.

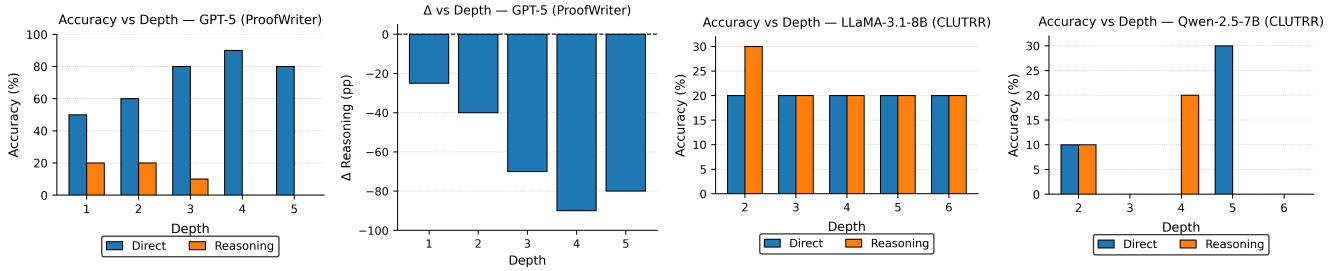

Figure 2: Accuracy vs. reasoning collapse. Comparison of direct and reasoning accuracy across depths for GPT-5 on ProofWriter, and LLaMA-3.1-8B and Qwen-2.5-7B on CLUTRR. Reasoning performance deteriorates as task depth increases, revealing distinct collapse points despite strong direct accuracy.

its reasoning accuracy drops quickly as complexity increases—here defined as the number of inference steps and the presence of compositional or logical operators required to reach the correct answer.

LLaMA-3.1-8B shows modest but steady gains, while Qwen-2.5-7B produces unstable spikes before collapsing. Error analysis points to recurring reasoning failures: in CLUTRR, depth-2 cases such as "shared parent $\Rightarrow$ siblings" are often mislabeled as parent–child, and in ProofWriter, negated entailments (e.g., "does not need") yield confident but inconsistent rationales. In many cases, models extended their explanations but reinforced early mistakes, suggesting that collapse reflects systematic weaknesses—especially with compositional rules and logical operators—rather than random noise.

Conclusion and Outlook

In this work, we evaluated the reasoning abilities of both closed- and open-source LLMs by systematically varying inference depth across CLUTRR and ProofWriter. Our framework reveals that reasoning prompts provide modest gains at shallow depths but collapse as complexity grows. Even GPT-5, which reaches 90% accuracy on ProofWriter at depth four in direct mode, shows negative deltas beyond depth two. These findings indicate that current benchmarks often underestimate depth-driven brittleness. By introducing balanced depth sampling, direct vs. reasoning prompts, and collapse depth as a metric, our framework offers a simple but effective robustness probe that exposes *where* reasoning fails, enabling more precise comparisons across models.

Future Work

Building on these findings, we plan to extend depth-graded evaluations beyond CLUTRR and ProofWriter to broader reasoning benchmarks, and to investigate training or architectural strategies that might improve collapse.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (RS-2024-00340882).

References

Shojaee*, P.; Mirzadeh*, I.; Alizadeh, K.; Horton, M.; Bengio, S.; and Farajtabar, M. 2025. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity.

Shojaee, P.; Mirzadeh, I.; Alizadeh, K.; Horton, M.; Bengio, S.; and Farajtabar, M. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Sinha, K.; Sodhani, S.; Dong, J.; Pineau, J.; and Hamilton, W. L. 2019. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tafjord, O.; and Clark, P. 2021. ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; Zhou, D.; et al. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.