Hierarchical Deconstruction of LLM Reasoning: A Graph-based Framework for Analyzing Knowledge Utilization

Miyoung Ko[1*], Sue Hyun Park[1*], Joonsuk Park[2,3,4†], Minjoon Seo[1†]

[1]KAIST AI, [2]NAVER AI Lab, [3]NAVER Cloud, [4]University of Richmond

[*]Equal contribution [†]Equal advising

EMNLP 2024

Link to arXiv

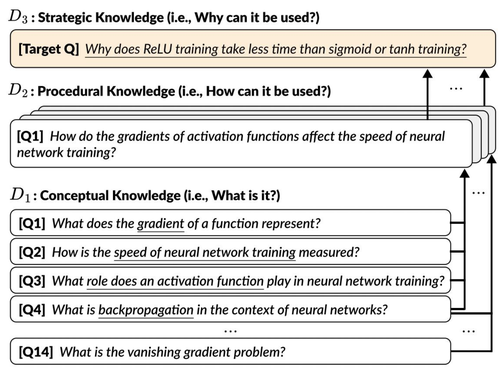
Introduction

Motivation: We aim to explore how large language models leverage multiple layers of knowledge to solve complex questions. By adopting Webb's Depth of Knowledge, we deconstruct real-world questions into hierarchical graphs across three levels of depth. This approach allows us to analyze performance discrepancies and uncover patterns related to model size and memorization.

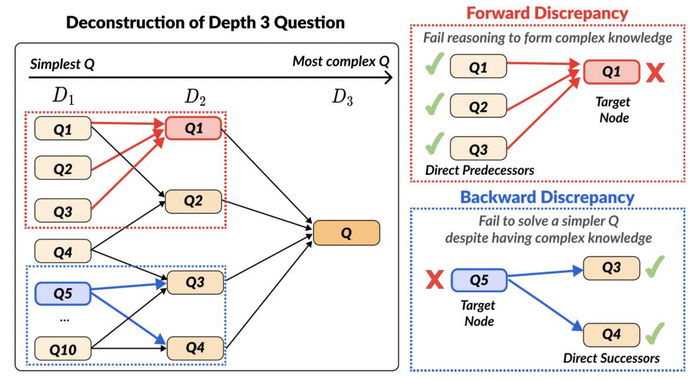
Contributions:

1) We propose to connect complex questions with simpler sub-questions by deconstructing questions based on depth of knowledge.
2) We design the DepthQA dataset. We measure forward and backward reasoning discrepancies across different levels of question complexity.
3) We investigate the reasoning abilities of LLMs with various capacities, analyzing the impact of model size and training data memorization on discrepancies. We demonstrate the benefits of structured, multi-turn interactions to perform complex reasoning.

Overview



[Figure 1] Example of reasoning across three levels of knowledge depth.



[Figure 2] Hierarchical structure of a deconstructed Depth 3 question, illustrating both forward and backward discrepancies.

DepthQA Dataset

Construction: Top-down deconstruction of D_3 questions

- From TutorEval (Chevalier et al., 2024), collect D_3 questions
- For each D_3 questions, deconstruct into D_2 questions using GPT-4 Turbo. (Same process for each D_2 to D_1)
- Post-processing based on three criteria; Comprehensiveness, Implicitness, Non-binary Questioning

Domain	# Questions			# Edges between questions	
	D_1	D_2	D_3	$D_1 \rightarrow D_2$	$D_2 \rightarrow D_3$
Math	573	193	49	774	196
Computer Science	163	54	14	212	55
Environmental Science	147	44	11	175	44
Physics	140	40	10	154	40
Life Sciences	98	28	7	111	28
Math $\rightarrow$ {CS, Physics}	-	-	-	11	0
Total	1,121	359	91	1,437	363

[Table 1] Statistics of DepthQA.

Depthwise Knowledge Reasoning Results

Average Accuracy [1,5]
: Average factual accuracy scored by LLM-as-a-judge

Discrepancy Metrics [0, 1]

$$\text{Forward Discrepancy}(q_k) = \max\left(0, \frac{1}{4}\left(\text{avg}_{q \in DP(q_k)}[f(q)] - f(q_k)\right)\right)$$

$$\text{Backward Discrepancy}(q_k) = \max\left(0, \frac{1}{4}\left(\text{avg}_{q \in DS(q_k)}[f(q)] - f(q_k)\right)\right)$$

Discrepancy = Intensity × Frequency

Fwd: ↑ ↓
Bwd: ↓ ↑

Model	Average Accuracy ↑				Forward Discrepancy ↓			Backward Discrepancy ↓		
	D_1	D_2	D_3	Overall	$D_2 \rightarrow D_3$	$D_1 \rightarrow D_2$	Overall	$D_2 \rightarrow D_3$	$D_1 \rightarrow D_2$	Overall
LLaMA 2 7B Chat	3.828	3.320	3.165	3.673	0.130	0.181	0.176	0.219	0.110	0.134
LLaMA 2 13B Chat	4.289	3.872	3.615	4.155	0.152	0.158	0.157	0.126	0.078	0.088
LLaMA 2 70B Chat	4.495	4.153	4.022	4.390	0.126	0.136	0.134	0.136	0.063	0.079
Mistral 7B Instruct v0.2	4.280	3.897	4.000	4.176	0.092	0.157	0.147	0.144	0.070	0.088
Mixtral 8x7B Instruct v0.1	4.599	4.532	4.429	4.574	0.087	0.079	0.081	0.063	0.063	0.063
LLaMA 3 8B Instruct	4.482	4.351	4.286	4.440	0.083	0.096	0.093	0.088	0.072	0.075
LLaMA 3 70B Instruct	**4.764**	**4.749**	**4.648**	**4.754**	**0.065**	**0.050**	**0.053**	**0.043**	**0.044**	**0.044**
GPT-3.5 Turbo	4.269	4.251	4.011	4.250	0.100	0.072	0.078	0.046	0.067	0.063

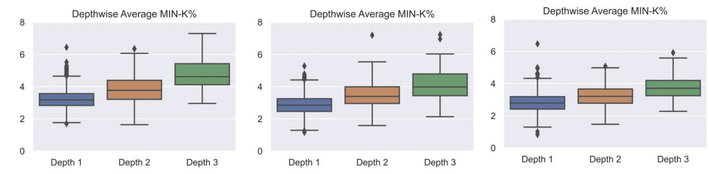
[Table 2] Depthwise reasoning performance of large language models.

Further Results

Depthwise Memorization:

Min-K% Probability: Average of negative log-likelihood of the K% least probable tokens.

- Higher Min-K% = Smaller possibility of memorization
- *Models rely less on memorization for complex questions* ($D_1 < D_2 < D_3$)



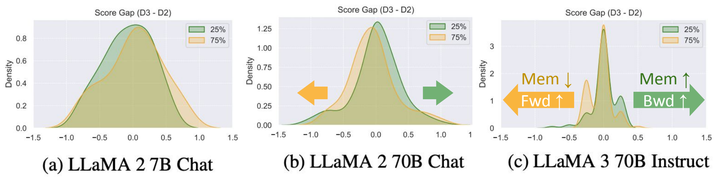
(a) LLaMA 2 7B Chat (b) LLaMA 2 70B Chat (c) LLaMA 3 70B Instruct

[Figure 3] Distribution of average Min-K%.

Memorization Gaps between Depths:

Memorization gap: [Accuracy(D_3) - Accuracy(D_2)] / 4

- Gap > 0: Backward discrepancy ($D_3 > D_2$)
- Gap < 0: Forward discrepancy ($D_3 < D_2$)
- For large models (70B), *potential causes of each discrepancy are different.*
- *Example Bwd: reason complex formula (D_3), fail to apply it (D_2)*



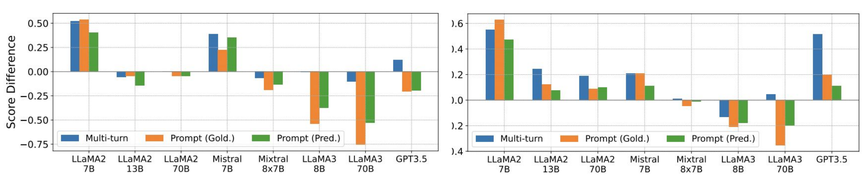
(a) LLaMA 2 7B Chat (b) LLaMA 2 70B Chat (c) LLaMA 3 70B Instruct

[Figure 4] Memorization gap for questions whose Min-K% probability is in the bottom 25% or top 75%.

Effect of Explicit Reasoning Process:

1) *Multi-turn*: Shallower Qs as multi-turn conversation
2) *Prompt (Gold)*: Shallower Qs + gold answers in prompts
3) *Prompt (Pred.)*: Shallower Qs + model's predictions in prompts

- *(Explicit) Shallower solutions are beneficial for small models and complex questions.*
- *(Implicit) Multi-turn interactions best improves performance.*



[Figure 5] Performance change after providing shallower questions