Information Re-Organization Improves Reasoning in Large Language Models

Xiaoxia Cheng, Zeqi Tan, Wei Xue, Weiming Lu†

College of Computer Science and Technology, Zhejiang University

{zjucxx, zqtan, lokilanka, luwm}@zju.edu.cn

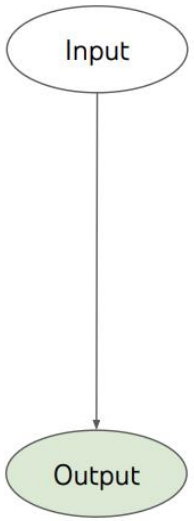
† Correspondence

Speaker: Xiaoxia Cheng

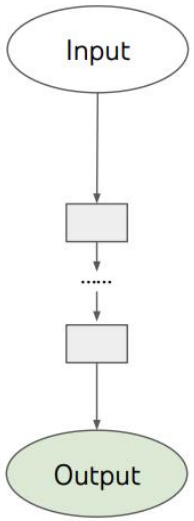
Existing Methods in Large Language Model Reasoning

Decompose the answer generation process into multiple intermediate steps.

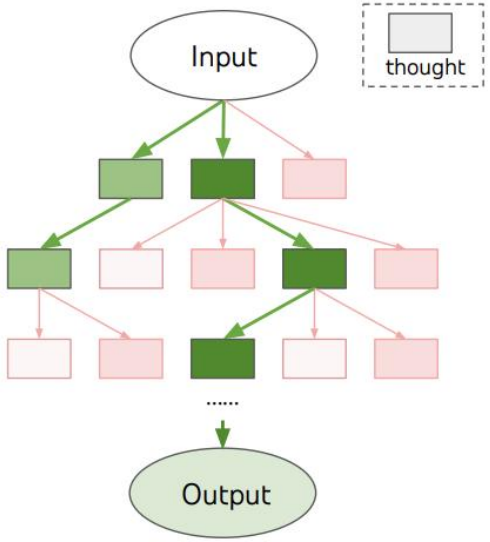
Direct



Chain of Thought(CoT)

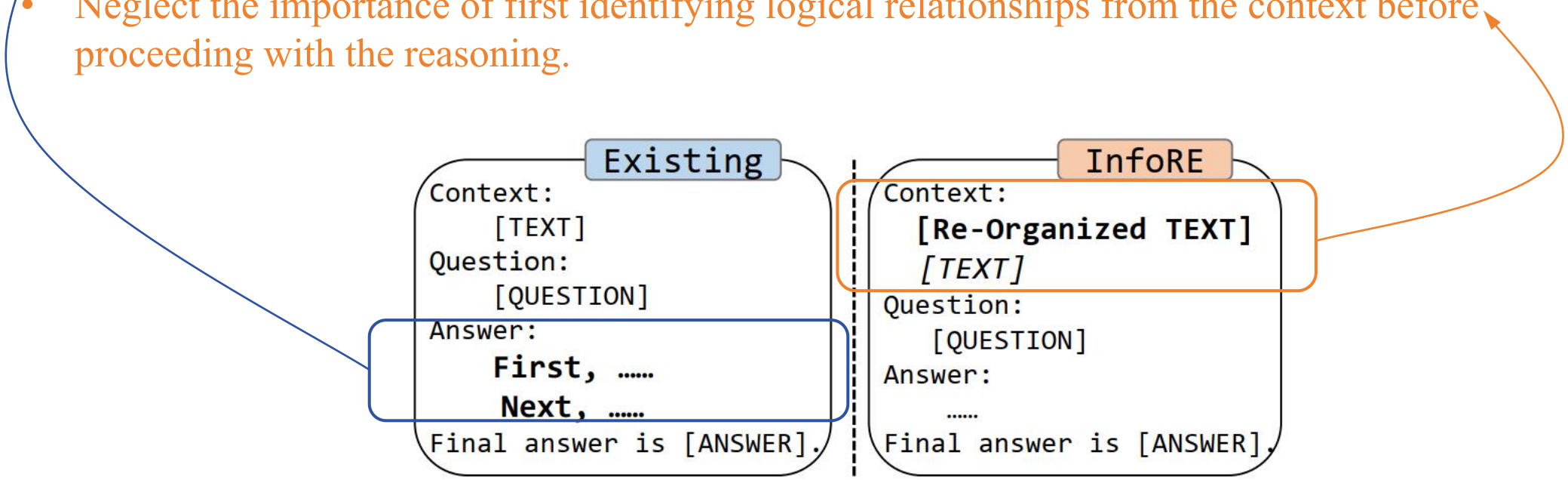


Tree of Thoughts(ToT)

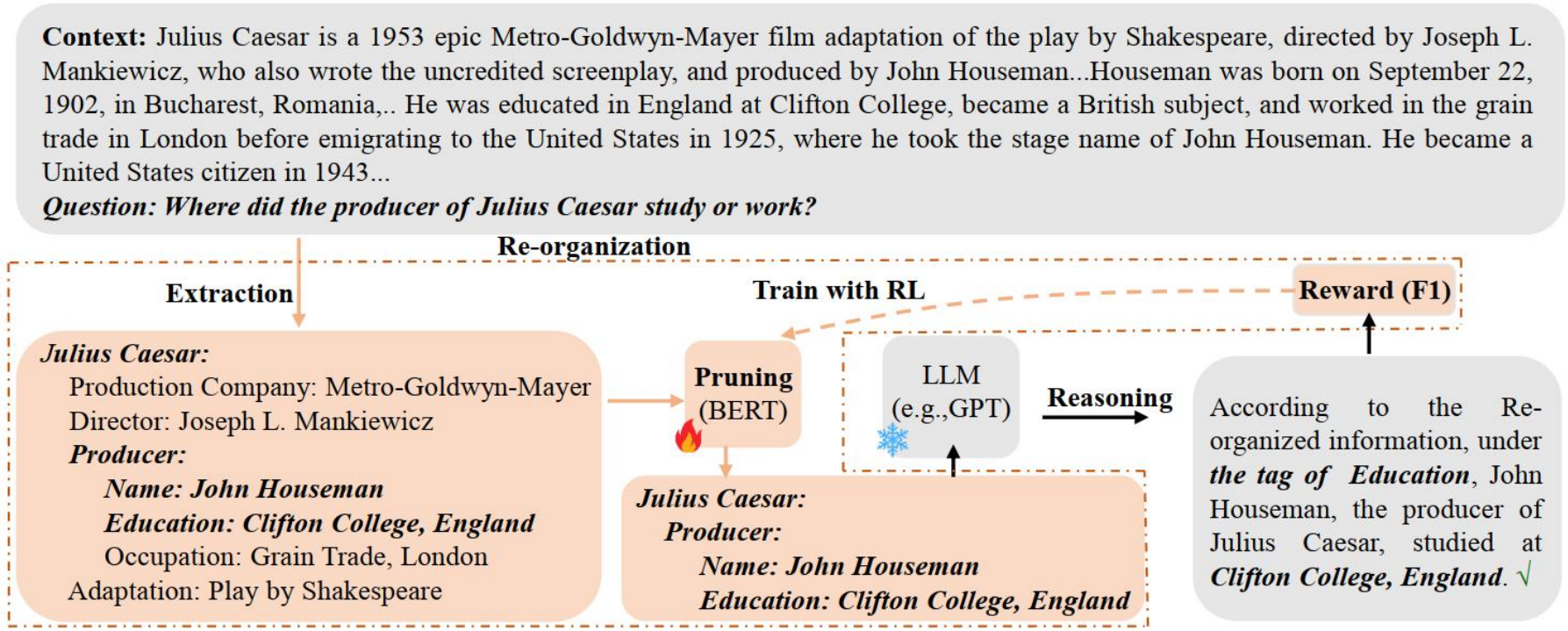


Challenge of Existing Reasoning Methods

- Focus on improving the reasoning process
- Neglect the importance of first identifying logical relationships from the context before proceeding with the reasoning.



Leverage Information Re-Organization Before Reasoning



We propose information re-organization (InfoRE) to improve reasoning , which consists two component:
-- Extraction to uncover the implicit logical relationships within the contextual content.
-- Pruning to further minimize noise that is irrelevant to the reasoning objective.

Extraction

The extraction uncovers the implicit logical relationships within the contextual content by transforming the content into a MindMap structure.

Context: Julius Caesar is a 1953 epic Metro-Goldwyn-Mayer film adaptation of the play by Shakespeare, directed by Joseph L. Mankiewicz, who also wrote the uncredited screenplay, and produced by John Houseman...Houseman was born on September 22, 1902, in Bucharest, Romania,.. He was educated in England at Clifton College, became a British subject, and worked in the grain trade in London before emigrating to the United States in 1925, where he took the stage name of John Houseman. He became a United States citizen in 1943...
Question: Where did the producer of Julius Caesar study or work?

LLM

Given a claim and corresponding evidence, please summarize the evidence as a mind map according to the claim. The output must be in a strict JSON format: {"mind_map": "mind_map"}.
CLAIM: [CLAIM]
EVIDENCE: [EVIDENCE]

Julius Caesar:
Production Company: Metro-Goldwyn-Mayer
Director: Joseph L. Mankiewicz
Producer:
Name: John Houseman
Education: Clifton College, England
Occupation: Grain Trade, London
Adaptation: Play by Shakespeare

Logical Relationship
e.g.: Director, Producer, Adaptation, ***Education*** Occupation

Multi-hop Connections:
e.g.: Julius Caesar → Producer → Education
Julius Caesar → Producer → Occupation

Pruning

After the extraction, not all logical relationships help answer the question. On the contrary, some may even interfere with the response to the question.

Context: Julius Caesar is a 1953 epic Metro-Goldwyn-Mayer film adaptation of the play by Shakespeare, directed by Joseph L. Mankiewicz, who also wrote the uncredited screenplay, and produced by John Houseman...Houseman was born on September 22, 1902, in Bucharest, Romania,.. He was educated in England at Clifton College, became a British subject, and worked in the grain trade in London before emigrating to the United States in 1925, where he took the stage name of John Houseman. He became a United States citizen in 1943...
Question: Where did the producer of Julius Caesar study or work?

LLM

Given a claim and corresponding evidence, please summarize the evidence as a mind map according to the claim. The output must be in a strict JSON format: {"mind_map": "mind_map"}.
CLAIM: [CLAIM]
EVIDENCE: [EVIDENCE]

Julius Caesar:
Production Company: Metro-Goldwyn-Mayer
Director: Joseph L. Mankiewicz
Producer:
Name: John Houseman
Education: Clifton College, England
Occupation: Grain Trade, London
Adaptation: Play by Shakespeare

distracting element

irrelevant to the question

Pruning

We designed a pruning operation to further reduce noise unrelated to the reasoning objective. This pruning model is based on the pre-trained model BERT and is fine-tuned using reinforcement learning.

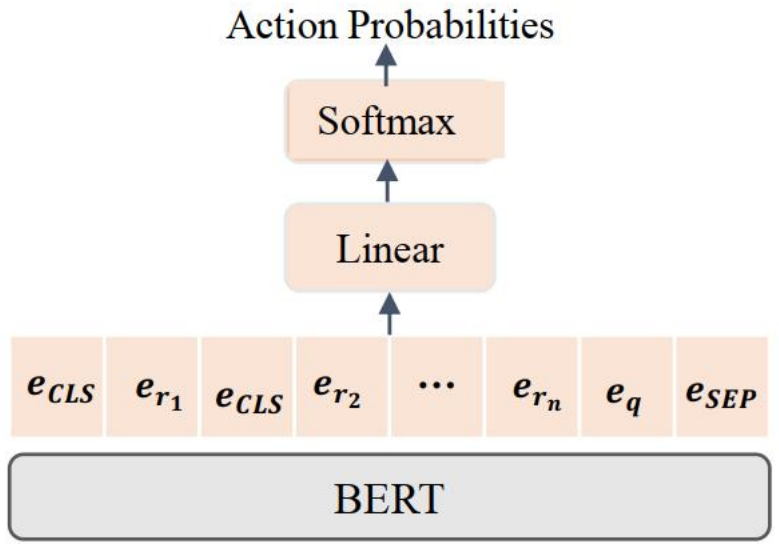


Figure 3: Illustration of Pruning model. The representation of [CLS] is used to obtain action probabilities.

$$\mathbb{E}_{\pi}[r] = \mathbb{E}_{\boldsymbol{g}\sim\mathcal{G},\boldsymbol{q}\sim\mathcal{Q},\boldsymbol{z}\sim\pi(\cdot|\boldsymbol{x},\boldsymbol{q})}[r(\boldsymbol{z},\boldsymbol{q})] \qquad (2)$$

$$r(\boldsymbol{z},\boldsymbol{q}) = \min(\mathcal{R}(\boldsymbol{z},\boldsymbol{q}), \text{clip}(\pi(\boldsymbol{z} \mid \boldsymbol{x},\boldsymbol{q}), 1-\epsilon, 1+\epsilon))$$

Experimental Results

Using only a zero-shot setting, our information re-organization (InfoRE) achieves an average absolute improvement of 4% across various contextually aware multi-hop reasoning tasks on Llama2, GPT-3.5, and GPT-4.

Table 1: Zero-shot performance on claim verification task across three large language models.

LLMs	Methods	HOVER			FEVEROUS	SCIFACT
		2-hop	3-hop	4-hop		
LLAMA2-70B	Standard	49.41	48.35	47.82	63.39	60.70
	InfoRE	**52.83**	**51.42**	**50.04**	**67.84**	**63.81**
		↑ 3.42	↑ 3.07	↑ 2.22	↑ 4.45	↑ 3.11
	CoT	50.02	48.76	48.01	64.53	61.24
	InfoRE + CoT	**53.20**	**51.70**	**50.15**	**68.12**	**64.02**
		↑ 3.18	↑ 2.94	↑ 2.14	↑ 3.59	↑ 2.78
GPT-3.5	Standard	64.74	63.04	61.54	87.67	77.42
	InfoRE	**68.21**	**66.45**	**64.91**	**91.31**	**81.54**
		↑ 3.47	↑ 3.41	↑ 3.37	↑ 3.64	↑ 4.12
	CoT	66.70	64.52	62.69	88.67	78.49
	InfoRE + CoT	**69.02**	**67.53**	**65.66**	**91.53**	**82.26**
		↑ 2.32	↑ 3.01	↑ 2.97	↑ 2.86	↑ 3.77
GPT-4	Standard	72.40	71.02	70.06	92.33	91.40
	InfoRE	**75.87**	**74.06**	**73.08**	**95.62**	**93.67**
		↑ 3.47	↑ 3.04	↑ 3.02	↑ 3.29	↑ 2.27
	CoT	73.82	72.07	70.68	92.67	92.47
	InfoRE + CoT	**76.69**	**75.16**	**73.62**	**95.67**	**94.32**
		↑ 2.87	↑ 3.09	↑ 2.94	↑ 3.00	↑ 1.85

Table 2: Zero-shot results on Question Answering and Reading Comprehension tasks. 2WMHQA, SQA, and HQA are abbreviations for 2WikiMultiHopQA, StrategyQA, and HotpotQA, respectively.

LLMs	Methods	2WMHQA	MuSiQue	SQA	HQA	WIKIHOP
LLAMA2 (70B)	Standard	52.56	49.55	51.23	66.07	40.32
	InfoRE	**57.62**	**52.78**	**55.32**	**69.98**	**42.90**
		↑ 5.06	↑ 3.23	↑ 4.09	↑ 3.91	↑ 2.58
	CoT	52.99	52.90	56.80	66.80	41.07
	InfoRE + CoT	**57.72**	**56.10**	**59.93**	**70.60**	**43.37**
		↑ 4.73	↑ 3.20	↑ 3.13	↑ 3.80	↑ 2.30
GPT-3.5	Standard	58.25	55.01	59.39	73.30	48.92
	InfoRE	**64.58**	**58.03**	**63.16**	**77.12**	**51.87**
		↑ 6.33	↑ 3.02	↑ 3.77	↑ 3.82	↑ 2.95
	CoT	59.37	57.05	67.51	73.90	49.65
	InfoRE + CoT	**65.13**	**60.52**	**70.45**	**77.74**	**52.70**
		↑ 5.76	↑ 3.47	↑ 2.94	↑ 3.84	↑ 3.05
GPT-4	Standard	72.69	62.65	68.32	79.33	55.46
	InfoRE	**76.52**	**66.36**	**71.20**	**83.22**	**58.01**
		↑ 3.83	↑ 3.71	↑ 2.88	↑ 3.89	↑ 2.55
	CoT	74.08	64.36	68.50	80.66	56.02
	InfoRE + CoT	**78.60**	**69.11**	**71.54**	**84.26**	**58.91**
		↑ 4.52	↑ 4.75	↑ 3.04	↑ 3.60	↑ 2.89

Analysis: Ablation Study and Effect of Re-organized Information Quality

Ablation Study

Both operations extraction and pruning included in information re-organization are essential.
RL-based pruning is more effective than similarity-based pruning.

Extraction is more effective than pruning.

Table 3: F1 performance of ablation studies.

Methods	2WikiMultiHopQA
Full model	**64.58**
w/o extraction	61.64 ↓2.94
w/o pruning	63.05 ↓1.53
similarity-based pruning	63.32 ↓1.26

Effect of Re-organized Information Quality

High-quality contextual re-organization enhances reasoning, while lower-quality re-organization, though less effective, still outperforms traditional methods without reorganization, demonstrating strong generalization.

Table 5: F1 performance of cross-validation, where InfoRE* denotes reason with GPT-3.5 but information re-organization with GPT-4, InfoRE† denotes reason with GPT-4 but information re-organization with GPT-3.5 (text-davinci-003).

Methods	FEVEROUS	2WikiMultiHopQA
GPT-3.5 (†)		
Standard	87.67	58.25
InfoRE	91.31 ↑1.19	64.58 ↑2.03
InfoRE *	**92.50**	**66.61**
GPT-4 (*)		
Standard	92.33	72.69
InfoRE	**95.62** ↓0.95	**76.52** ↓1.45
InfoRE†	94.67	75.07

Analysis: Error Analysis

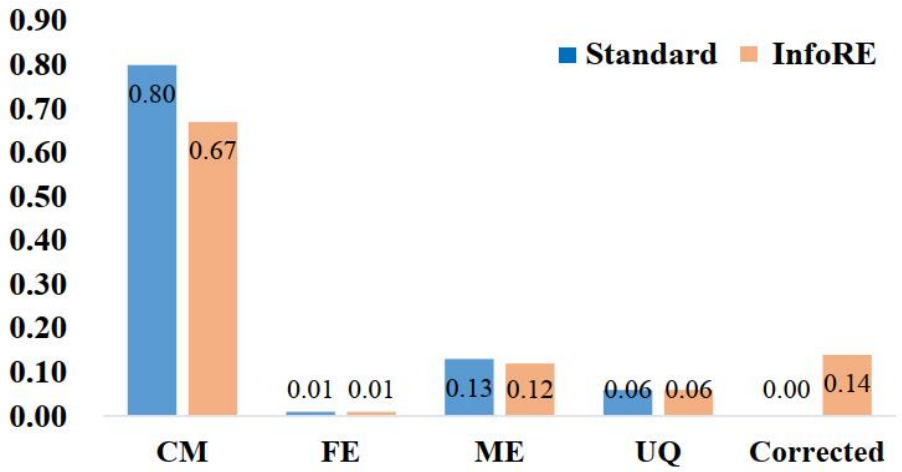


Figure 4: Error Analysis of InfoRE on 2WikiMultiHopQA against Standard baseline method. The first four rectangles are error categories, while "Corrected" on the far right denotes the percentage of errors originally made by the baseline method that our method InfoRE has successfully corrected.

- **Contextual Misunderstanding (CM):** This happens when the model fails to interpret or connect multiple pieces of information from different parts of the documents. Multi-hop reasoning requires synthesizing information from various segments, and recognizing logical relations, and any misunderstanding can lead to incorrect conclusion.
- **Factual Error (FE):** The model may provide an answer that is factually incorrect or not supported by the given documents. This is often due to the model's reliance on its training data, which may not always align with the specific facts in the context.
- **Mathematical Error (ME):** The error occurs when math calculations are involved in deriving the final answer.
- **Unanswerable Question (UQ):** It's a specific type of error or limitation in dataset design, where the context does not contain enough information to provide a valid answer to the posed question.

- Among the four types of error, contextual misunderstanding is the primary source of errors in the baseline.
- Our method (InfoRE) mainly corrects 14% of errors coming from the baseline method, most of the corrected errors are contextual misunderstanding errors.

Conclusion

To tackle the challenge of existing works neglecting the deeper understanding of contex before reasoning, we propose an information re-organization method to improve the reasoning ability of LLMs.

The information re-organization method

- First uncovers the logical relationships, multi-hop connections of context with an extraction component.
- Then prune the irrelevant information with a pruning component.
- Experimet improvements on contextually aware multi-hop reasoning tasks on LLAMA, ChatGPT, GPT-4 show the efficacy of our proposed method.

Despite the effectiveness of our method, the extraction operation relies on LLMs. In the future, we will explore more about how to integrate small language model for logical relationship extraction.

浙江大學 ZHEJIANG UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks for your listening!

Speaker: Xiaoxia Cheng