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

Abstract

Improving the reasoning capabilities of large language models (LLMs) has attracted considerable interest. Recent approaches primarily focus on improving the reasoning process to yield a more precise final answer. However, in scenarios involving contextually aware reasoning, these methods neglect the importance of first identifying logical relationships from the context before proceeding with the reasoning. This oversight could lead to a superficial understanding and interaction with the context, potentially undermining the quality and reliability of the reasoning outcomes. In this paper, we propose an information re-organization (**InfoRE**) method before proceeding with the reasoning to enhance the reasoning ability of LLMs. Our re-organization method involves initially extracting logical relationships from the contextual content, such as documents or paragraphs, and subsequently pruning redundant content to minimize noise. Then, we utilize the re-organized information in the reasoning process. This enables LLMs to deeply understand the contextual content by clearly perceiving these logical relationships, while also ensuring high-quality responses by eliminating potential noise. To demonstrate the effectiveness of our approach in improving the reasoning ability, we conduct experiments using Llama2-70B, GPT-3.5, and GPT-4 on various contextually aware multi-hop reasoning tasks. Using only a zero-shot setting, our method achieves an average absolute improvement of 4% across all tasks, highlighting its potential to improve the reasoning performance of LLMs. Our source code is available at `https://github.com/hustcxx/InfoRE`.

1 Introduction

Large language models (LLMs) demonstrate powerful generative capabilities and achieve remarkable performance across a range of linguistic tasks [1–3]. However, their capabilities in performing complex reasoning tasks — an essential aspect of advanced language understanding and intelligent decision-making — still present substantial challenges [4, 5]. This has spurred researchers to explore innovative strategies [6–8] to improve the reasoning capabilities of these models.

Recently, diverse methods have been developed to enhance the reasoning ability of LLMs. For example, a notable method Chain-of-Thought (CoT) [6], incorporates a series of intermediate reasoning steps into the reasoning. CoT [6] allows for a more transparent and understandable path to the final answer, making it easier to follow the logic behind the conclusion. Building upon this foundation, subsequent approaches such as Tree of Thoughts (ToT) [7] and Graph of Thoughts (GoT) [8] are proposed to further refine the reasoning steps and enhance the accuracy and reliability of LLMs. Different from the sequential intermediate steps of CoT [6], ToT [7] and GoT [8] model the problem solving process into structured tools of tree and graph, respectively. A critical observation

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-4137

is that these existing approaches primarily focus on improving the reasoning process of LLMs, as shown in Figure 1 (left). However, in scenarios involving contextually aware reasoning, it is equally important to first identify logical relationships of context before proceeding with the reasoning, not just improve their reasoning process. This is because logical relationships, such as parallelism, causal connections, contrasts, etc., are essential elements of reasoning [9]. Nevertheless, these existing methods often neglect this crucial step. Such an oversight can lead to a superficial understanding and interaction with the context, potentially undermining the quality of the reasoning results.

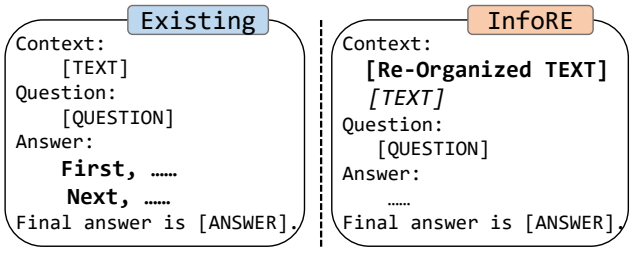

Figure 1: InfoRE (Ours) vs existing methods. In contrast to the existing methods that primarily focus on the reasoning process, our InfoRE emphasizes the re-organization of context information. The *[TEXT]* in italics indicates that it is optional in the reasoning process.

Inspired by the fact that when faced with context-aware reasoning tasks humans often first re-organize existing contextual information to uncover the logical relationships, eliminate noises, and enhance their understanding of the context, we propose an information re-organization (**InfoRE**) method, to ground reasoning by the re-organized information. As shown in Figure 1 (right), different from previous methodologies that primarily focus on refining the reasoning steps to enhance the reasoning capabilities of LLMs, our approach takes a novel direction of context re-organization. We emphasize the utilization of re-organized contextual content to explicitly present the logical relationships that are often implicit within the plain text, promoting more effective reasoning. Specifically, our re-organization method comprises two operations: extraction and pruning. The extraction first uncovers the implicit logical relationships within the contextual content by transforming the content into a MindMap structure [10]. We employ this structure because it is rich in logical relationships and encompasses multi-hop connections. Pruning is then used to further minimize noise that is irrelevant to the reasoning objective. The pruning operation uses a pre-trained BERT [11] based model trained with reinforcement learning (RL). Finally, we utilize the re-organized context to reason. This enables LLMs to deeply understand the context by clearly perceiving these logical relationships, facilitating the quality and reliability of reasoning. Besides, our information re-organization method can be integrated with existing prompt methods, like CoT [6], to further improve the reasoning ability of LLMs. To verify the efficacy of our proposed InfoRE method, we conduct experiments using Llama2-70B [2], GPT-3.5 [1], and GPT-4 [3] on various contextually aware multi-hop reasoning tasks, including claim verification [12], question answering [13], and reading comprehension [14]. Using only a zero-shot setting, our method achieves an average improvement of 4% across all tasks, highlighting its potential to improve the reasoning performance of LLMs.

Our main contributions are as follows:

- In contrast to existing methods that primarily focus on refining the reasoning steps to enhance the reasoning capabilities of LLMs, we take a novel direction of context re-organization.
- The re-organization method initially uncovers the logical relationships that encompass multi-hop connections in the contextual content by extraction, and subsequently minimizes noise by pruning.
- Experiment improvements on contextually aware multi-hop reasoning tasks across claim verification, question answering, and reading comprehension show the efficacy of our proposed method.

2 Related Work

Reasoning with LLMs LLMs [1–3] have revolutionized the field of natural language processing (NLP) and demonstrate remarkable proficiency across a range of linguistic tasks. To further improve

the reasoning ability of LLMs has attracted considerable interest. In-context learning [1], as a promising approach to enhance the reasoning abilities of LLMs, has been verified in mathematical reasoning task [6]. In addition, CoT [6] incorporates a coherent series of intermediate reasoning steps to improve the reasoning ability of LLMs. Following this paradigm, ToT [7] and GoT [8] have also been proposed to focus on improving the structure of the intermediate chain. Then self-consistency [15] and plan-to-solve [15] focus on the reliability of the chain. Recently, the step-back prompting [16] is proposed, which obtains the high-level concept and first principles from instances by abstraction in the first step, then guides the reasoning of LLMs with the obtained concept and principles. For tasks that require multi-hop reasoning, current approaches [17, 18] generally decompose multi-hop problems into simpler sub-tasks problems relying on the in-context learning ability of LLMs. In contrast to existing methods, we take a novel direction of context re-organization to enhance the reasoning capabilities of LLMs for contextually aware reasoning tasks.

Information Re-organization Information re-organization is a technique that leverages other structures to improve the clarity and comprehensibility of information. Moreover, it can also reveal conceptual relationships implicit in the original textual text, making it a valuable strategy for various tasks. For example, SG [19] re-organizes the document to scattered knowledge graph triples for explainable multi-hop question answering. GraphRAG [20] build a community hierarchy and generate summaries for these communities after extracting a knowledge graph, then leverage these structures on RAG-based tasks. MindMap [10] is a powerful structure method for representing knowledge and concepts, it can be used to construct a hierarchical abstraction of natural language text [21]. Previous method [22] uses it to organize a large amount of scientific material to enhance the clarity of the text material. Different from the scattered knowledge graph triples used in SG [19], MindMap is more centralized, which aggregates content related to the same topic together. In this paper, we use it as the structure of re-organized information to uncover the logical relationships and multi-hop connections implicit within the plain context with one step, as opposed to GraphRAG [20] separately aggregates information with multiple steps.

3 Methodology

In this section, we first give a formulation of a context-aware reasoning task in §3.1 and then describe our method in detail. As shown in Figure 2, the framework of our method consists of two components, an information re-organization §3.2 and a reasoning step using the re-organized context §3.3.

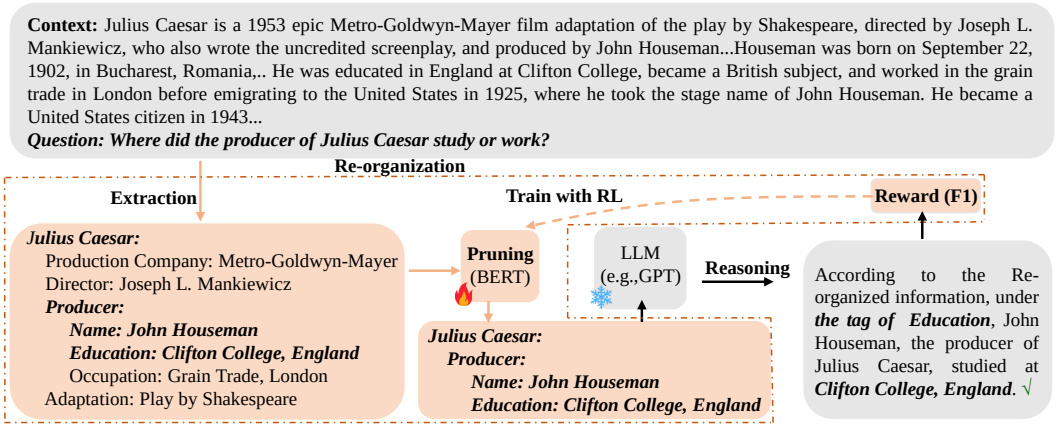

Figure 2: Illustration of our information re-organization method with two modules: 1)Information Re-Organization, which includes logic relationship extraction and noise pruning. 2) Reasoning using re-organized context. The re-organized context in ***black italicized*** text is relevant to the question.

3.1 Task Formulation

Given a sample (c, q, a) from sample space $(\mathcal{C}, \mathcal{Q}, \mathcal{A})$, where c denotes gold context, q denotes question, the task aims to obtain answer a to the question q, based on gold context c with a large language model. In particular, it requires multi-hop reasoning to get the answer a to question q.

https://doi.org/10.52202/079017-4137

3.2 Information Re-Organization

The purpose of the information re-organization is to obtain the logical relationships from the context and minimize the noise irrelevant to the question. This goal is achieved through two operations: extraction and pruning.

3.2.1 Extraction

For a question q, if the derivation of its answer a relies on context c, then a deeper understanding of c is crucial. Logical relationships, such as parallelism, causal connections, contrasts, etc., are essential elements of understanding and reasoning [9]. However, the logical relationships in the plain context are often implicit. Therefore, we perform a extraction operation on the plain context to uncover the logical relations in it using a language model. The process can be defined as:

$$g = f_\theta(c, q, P_{in}) \tag{1}$$

where $f_\theta(\cdot)$ is a language model parameterized by θ, P_{in} is input task prompt. The specific task prompt P_{in} used in our paper is displayed in the Appendix A. We perform Equation 1 for each c in the sample space to obtain the extracted context space $\mathcal{G}$. We use the MindMap [10] structure to display the reorganized content, because it serves as a powerful structure for representing knowledge, concepts, and perspectives, contains not only logical relationships but also multi-hop connections.

As shown in Figure 2 (bottom left), the extracted context $g \in \mathcal{G}$ contains not only parallel logical relationships, e.g., Director&Producer, but also causal relationships, e.g., Julius Caesar→Director. Furthermore, it also describes the three-hop connections, such as Julius Caesar→Director→Name, Julius Caesar→ Director→Occupation, and Julius Caesar→Director→Education. This information with logical relationships enables LLMs to deeply understand the contextual content by clearly perceiving these logical relationships, facilitating the quality and reliability of reasoning. Additionally, this multi-hop connection corresponds to the complex multi-hop problem and therefore helps to solve this multi-hop question q.

3.2.2 Pruning

As described in Section 3.2, the extracted context $g \in \mathcal{G}$ contains various logical relationships and rich attributes. However, not all logical relationships and attributes help answer the question q. On the contrary, some may even interfere with the response to the question. For example, consider the question in Figure 2, the content "Julius Caesar → Production Company" is a distracting element, and "Julius Caesar → Adaptation" is irrelevant to the question.

To further reduce the interference of distracting or irrelevant logical relations and attributes on the retrieval of answers for question q, we use a pruning model trained through reinforcement learning (RL). The pruning model is based on the pre-trained BERT [11] due to its high generalizability, as shown in Figure 3. Its input consists of concatenated logical relationships, their corresponding attribute values from g, and the question q. For example, to prune the relation Julius Caesar → Adaptation, the input is: [CLS] Julius Caesar Adaptation Play by Shakespeare [SEP] Question [SEP]. We give a detailed demonstration of input format in Appendix B.

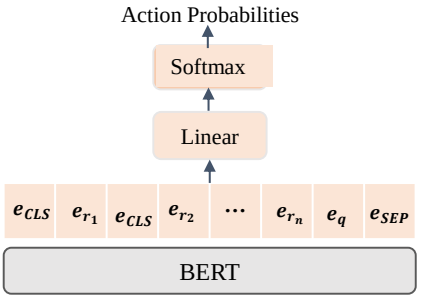

Figure 3: Illustration of Pruning model. The representation of [CLS] is used to obtain action probabilities.

RL Formulation We formulate the pruning policy model optimization as an RL problem and employ proximal policy optimization (PPO) [23]. The action is keeping or deleting the logical relations. The policy decides the action probability given the question and g. We fine-tune the policy model π by optimizing the reward r:

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

Reward Function Our goal is to maximize LLM's generation toward the desired target by an alignment measure $\mathcal{R}$, and we use this as a reward. In this paper, the alignment metric we have

chosen is the F1 score. To keep the policy network π from moving too far from the old result, we add a clipped term in the reward. Therefore, the final reward becomes:

$$r(\boldsymbol{z}, \boldsymbol{q}) = \min(\mathcal{R}(\boldsymbol{z}, \boldsymbol{q}), \text{clip}(\pi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{q}), 1-\epsilon, 1+\epsilon)) \tag{3}$$

where ϵ is a hyperparameter indicating the range of CLIP to be performed. After pruning, the extracted contextual content g becomes the context $g^{'}$, which is closely related to the reasoning objective q. We perform a pruning operation for each $g \in \mathcal{G}$ to obtain the pruned context space $\mathcal{G}^{'}$.

3.3 Reasoning

After the re-organization process in Section 3.2, we get the re-organized context $g^{'} \in \mathcal{G}^{'}$. Then the re-organized context $g^{'}$ can be used as a context alone or combined with the original context c to get the final answer of q. The reasoning process can be defined as:

$$o = f_\theta(g^{'}, [c], q, P_r) \tag{4}$$

where P_r denotes prompt, and the content within [] denotes that it is optional. The specific prompt used in reasoning is displayed in the Appendix C.

4 Experimental

4.1 Tasks and Datasets

To verify the effectiveness of our information re-organization method, we conduct experiments across a range of contextually aware multi-hop reasoning tasks and datasets, including claim verification [12], question answering [13], and reading comprehension [14]. The detailed dataset information, including data splits and statistics, is available in Appendix D.

Claim Verification The task involves assessing a given claim against a set of evidence documents to determine whether they support or refute the claim [12]. We consider HOVER [24] and FEVEROUS [25], which comprise complex claims that necessitate multi-hop reasoning for verification. Besides, we also take into account the SCIFACT [26] dataset, notable for its inclusion of scientific claims.

Question Answering For this task, we consider the following datasets: 2WikiMultiHopQA [27], StrategyQA [28], MuSiQue [29], and HotpotQA [30]. To answer the questions in these datasets requires not only multi-hop reasoning but also cross-document analysis.

Reading Comprehension Machine reading comprehension task requires a model to process documents and select an answer from the provided candidates to a question about the content [14]. We primarily consider WIKIHOP [31] in the task, which necessitates multi-hop reasoning to derive the final answer. Additionally, HotpotQA [30] is also frequently considered as part of the QA domain.

4.2 Baselines

In our paper, we compare our method InfoRE to two reasoning baselines: 1) **Standard**. 2) **CoT**. The Standard approach is a method that directly reasons with the original textual context. The CoT [6] method involves augmenting standard reasoning methods by adding a step-by-step thought process. In our paper, we conduct the CoT strategy by appending the sentence "Let's think step by step." at the end of the question.

The baseline methods and our InfoRE both adopt a zero-shot setting to counteract the potential randomness associated with demonstrations in a few-shot setting. We also design an answer-format instruction within the prompts for various tasks to standardize the structure of the final answer, thereby enhancing the precision of answer extraction. Moreover, all results reported in the paper use only the reorganized contextual information to reason. Comprehensive details about prompts and answer-format instruction are available in Appendix C. Following previous methods [17, 18, 25], we run the official evaluation scripts of each dataset to get the F1 to measure the results.

4.3 Implementation Details

In our paper, the LLMs employed in the extraction and reasoning process include Llama2-70B [2], GPT-3.5 (text-davinci-003) [32] and GPT-4 [3]. We use the official version of Llama2-70B. The

 https://doi.org/10.52202/079017-4137

specific version of GPT-4 is GPT-4-0613. We configure all models with top_p parameter as 1.0 and temperature as 0.0. In the policy model, we use the BERT-base version on all tasks and datasets. In RL training, we calculate the F1 score between the generated answer and the reference answer as the reward, with a rescaling coefficient of 10. We train the model for 1000 episodes. We conduct training for epoch 5, a batch size of 4, and a learning rate of 2e-6. The parameter of ϵ is set to 0.2. All experiments are conducted on an NVIDIA RTX A6000.

5 Results and Analysis

5.1 Main Results

Claim Verification Table 1 presents a comprehensive performance comparison between our InfoRE and existing zero-shot techniques. For the HOVER dataset, we segment it into the 2-hop, 3-hop, and 4-hop levels following previous methods [17]. As depicted in Table 1, our InfoRE demonstrates significant improvements in the zero-shot claim verification task. The CoT [6] approach offers a lightly increase of 0.62% in the HOVER 4-hop using GPT-4, which indicates its marginal utility in more complex reasoning scenarios. Yet, our InfoRE achieves 3.02% improvement on the HOVER 4-hop using GPT-4 showing remarkable performance on contextual aware understanding and reasoning. This improvement is further increased to 73.62% in combination with CoT, suggesting that the methods complement each other effectively. In the case of Llama2-70B, the combined application of InfoRE and CoT yields a score of 53.20% on the 2-hop HOVER task, surpassing its CoT-only score of 50.02%. This pattern of improvement is consistent with GPT-3.5. GPT-4 shows superior performance across all methods and datasets, suggesting an inherent advanced reasoning ability. This performance is mirrored in the specialized benchmarks, where GPT-4 with InfoRE attains near-perfect accuracy on FEVEROUS (95.62%) and very high accuracy on SCIFACT (93.67%). These figures solidify the notion that the InfoRE indeed enhances the reasoning capabilities of LLMs.

Table 1: Zero-shot performance on claim verification task across three LLMs.

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

Question Answer and Reading Comprehension Different from claim verification task, this task involves using multiple documents as context, presenting a challenge in cross-document reasoning. Intuitively, when the reasoning process involves multiple documents, information re-organization can effectively merge the information from different documents and uncover logical relationships that are not apparent in plain text. Performance in Table 2 verifies the intuitive. We can see that after applying information re-organization, all the results have a significant improvement. GPT-4 outperforms

the other models, with the highest F1 score of 76.52%, 71.20%, and 83.22% when employing InfoRE on 2WikiMultiHopQA, StrategyQA, and HotpotQA, respectively. This is consistent with the performance improvements observed with GPT3.5 (text-davinci-003) and Llama2-70B when applying InfoRE. Different from conventional QA tasks, reading comprehension tasks require LLMs to not only deeply understand the context but also identify distractors among the candidates, increasing the reasoning challenge. Our InfoRE consistently shows improvements on this task. Specifically, in the WIKIHOP dataset, GPT-3.5 with InfoRE outperforms other methods with an F1 of 51.87%. This improvement further verifies the effectiveness of our method.

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

5.2 Analysis

Ablation studies In our paper, the re-organization comprises two components: extraction and pruning. To investigate the impact of each component in detail, we conduct a series of ablation experiments using GPT-3.5 on the 2WikiMultiHopQA dataset. First, we directly remove extraction and pruning from our method, and the results are shown in the second and third rows of Table 3, respectively. It is worth mentioning that in the experiment where extraction is removed, we directly prune the sentences in the original context. Furthermore, we replace the reinforcement learning-based pruning method with a similarity-based pruning method to demonstrate its effectiveness. Specifically, the similarity-based pruning method uses Siamese-BERT, which takes the original question and each logical relationship as inputs separately, and then generates the corresponding representations. Then, we calculate the cosine similarity between these two representations. Finally, we removed the 30% of logical relationships with the lowest similarity, the result is shown in the last row of Table 3.

The results in Table 3 show that removing the extraction and pruning operations leads to performance drops of 2.94% and 1.53%, respectively. This demonstrates the effectiveness of both components in our methods. The larger performance drop after removing extraction highlights the importance of extracting logical relationships for effective reasoning. After replacing the pruning model, the performance dropped by 1.26%, but the results were still better than without pruning.

Table 3: F1 performance of ablation studies.

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

https://doi.org/10.52202/079017-4137

This not only demonstrates the necessity of pruning but also highlights the effectiveness of the reinforcement learning-based pruning method.

Quality of Re-Organized Information To assess the quality of the re-organized information, we perform a quantitative evaluation of the re-organized information with GPT-4 (gpt-4-32k) on the 2WikiMultiHopQA dataset. Specifically, we select 100 samples from the dataset, GPT-4 (gpt-4-32k) is asked to rank re-organized information produced by GPT-3.5 (text-davinci-003) [1] and GPT-4, as well as original context following criteria: (1) Depth: The information present multiple relationships of a topic, offering insightful perspectives or in-depth understanding of the subject. (2) Clarity: Information is clear and precise, making it easy to understand without ambiguity. All of the information is ranked 1, 2, and 3 with 3, 2, and 1 scores, respectively. Finally, we get a weighted average score for each information to measure the overall quality.

The results in Table 4 demonstrate that the re-organized information outperforms the original textual context in terms of depth and clarity, which justifies the motivation of our paper. In addition, the re-organized information from GPT-4 outperforms GPT-3.5, which proves in another way that GPT-4 is indeed more capable than GPT-3.5. The evaluation results further validate the effectiveness of our method. Furthermore, the 22.22% improvement in depth is more obvious than the 15.14% improvement in clarity, which indicates that the re-organization of the information has been particularly effective in enhancing the logical relationships of the information. The improvement in clarity suggests that the information is now presented in a more direct and streamlined manner, making it easier for LLMs to grasp the essential points without wading through unnecessary details.

Table 4: Qualitative evaluation results on 2WikiMultiHopQA dataset. Avg R denotes the weighted average ranking score. The larger ranking score denotes better information quality.

Methods	**Depth**			
	1st	**2nd**	**3rd**	**Avg R.**
Original	0.22	0.36	0.42	1.80
GPT-3.5	0.32	0.36	0.32	2.00
GPT-4	0.46	0.28	0.26	**2.20**
Methods	**Clarity**			
	1st	**2nd**	**3rd**	**Avg R.**
Original	0.25	0.35	0.40	1.85
GPT-3.5	0.35	0.32	0.33	2.02
GPT-4	0.40	0.33	0.27	**2.13**

Effect of Re-organized Information Quality
To explore the effects of the quality of re-organized context on model reasoning capabilities, we utilize both GPT-3.5 (text-davinci-003) and GPT-4 to re-organize context information. Reasoning processes are then independently executed on each model. Moreover, employing this cross-validation technique also allows us to effectively evaluate the robustness of our method.

The cross-validation results, shown in Table 5, indicate that the use of either GPT-3.5 or GPT-4 for information re-organization leads to a marked improvement in the performance of LLMs on reasoning tasks. Additionally, it is observed that using GPT-4 for information re-organization while reasoning with the GPT-3.5 results in a further increase in reason results by 1.19% and 2.03% on FEVEROUS and 2WikiMultiHopQA datasets, respectively. This implies that enhancing the quality of re-organized information can improve the performance of LLMs, pointing to a promising direction for future research in refining information synthesis and re-organization. Conversely, when GPT-3.5 is employed for information re-organization in conjunction with reasoning using the GPT-4 model, there is a decrease in reason ability by 0.95% and 1.45% respectively. In addition, the magnitude of the decrease is lower than the increase, which implies that our re-organization strategy may have a greater impact on models with

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	64.58
InfoRE *	**92.50**	**66.61**
GPT-4 (*)		
Standard	92.33	72.69
InfoRE	**95.62**	**76.52**
InfoRE†	94.67	75.07

weaker reasoning abilities. This finding aligns with our understanding that models with inherently weaker reasoning abilities tend to rely more heavily on external strategy. For models with stronger inherent capabilities, our method still further improve its reasoning ability.

Error Analysis To better comprehend where the errors in our InfoRE methodology come from and where they are fixed, we annotate 100 wrong predictions made by both InfoRE and Standard methods with GPT-3.5 on 2WikiMultiHopQA dataset. We categorize the errors into 4 classes:

- **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.

When engaging in reasoning with LLMs, all four error categories are present. As shown in Figure 4, there are 6% unanswerable question errors in the dataset, and more than 90% errors occur in the reasoning in the baseline method. Among the four types of error, contextual misunderstanding is the primary source of errors in the baseline, highlighting the importance of an in-depth understanding of context in solving reasoning tasks with LLMs. This finding is consistent with our motivation presented in the introduction section. Moreover, comparing the results of errors between our method and the baseline method, our method mainly corrects 14% of errors coming from the baseline method, most of the corrected errors are contextual misunderstanding errors. This further indicates that our InfoRE method assists LLMs in understanding context, signifying the necessity and effectiveness of conducting information re-organization before directly addressing the original question.

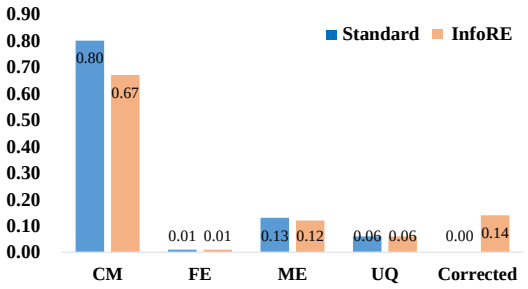

Figure 4: Error Analysis of InfoRE on 2WikiMultiHopQA against Standard baseline method. The first four rectangles are error categories, while "Corrected" on the far right denotes the percentage of errors originally made by the baseline method that our method InfoRE has successfully corrected.

6 Conclusion

In this paper, we propose an information re-organization method to improve the reasoning ability of LLMs. Compared with previous approaches primarily focus on improving the quality of intermediate steps, our method emphasizes uncovering the logical relationships, multi-hop connections, and pruning the irrelevant information through information re-organization. This approach enables LLMs to explicitly perceive the logical relationships and multi-hop connections of concepts within the context, promoting a deeper integration and understanding of the context, which results in more robust reasoning outcomes. To verify the effectiveness of our method, we conduct experiments using various LLMs across a range of contextually aware multi-hop reasoning tasks. The experiment results demonstrate the potential of our method to improve the reasoning ability of LLMs. Additionally, our method has a positive impact on various tasks involving context understanding, such as academic research, legal analysis, and medical diagnostics. However, it is also important to be aware of potential negative impacts, such as the propagation of misinformation.

https://doi.org/10.52202/079017-4137

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (No. 62376245), the "Pioneer" and "Leading Goose" R&D Programs of Zhejiang (No. 2024C03255), the Fundamental Research Funds for the Central Universities (226-2024-00170), the project of the Donghai Laboratory (Grant no. DH-2022ZY0013), National Key Research and Development Project of China (No. 2018AAA0101900), and MOE Engineering Research Center of Digital Library.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[2] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL http://arxiv.org/abs/2307.09288.

[3] OpenAI. Gpt-4 technical report, 2024. URL http://arxiv.org/abs/2303.08774.

[4] Konstantine Arkoudas. Gpt-4 can't reason, 2023. URL http://arxiv.org/abs/2308.03762.

[5] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 22–31, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701979. doi: 10.1145/3594536.3595163. URL https://doi.org/10.1145/3594536.3595163.

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[7] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models, 2023.

[8] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michał Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, Mar 2024. doi: 10.1609/aaai.v38i16.29720. URL https://ojs.aaai.org/index.php/AAAI/article/view/29720.

[9] M. D. S. Braine. On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, 85:1–21, 1978. doi: 10.1037/0033-295x.85.1.1.

[10] Tony Buzan, Barry Buzan, and James Harrison. The mind map book: Unlock your creativity, boost your memory, change your life. 2010. URL https://api.semanticscholar.org/CorpusID:141986951.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[12] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022. doi: 10.1162/tacl_a_00454. URL https://aclanthology.org/2022.tacl-1.11.

[13] Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. Unsupervised question decomposition for question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8864–8880, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.713. URL https://aclanthology.org/2020.emnlp-main.713.

[14] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246.

[15] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 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, 2023. URL https://openreview.net/pdf?id=1PL1NIMMrw.

[16] Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3bq3jsvcQ1.

[17] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.386. URL https://aclanthology.org/2023.acl-long.386.

[18] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Question decomposition improves the faithfulness of model-generated reasoning, 2023. URL http://arxiv.org/abs/2307.11768.

[19] Ruosen Li and Xinya Du. Leveraging structured information for explainable multi-hop question answering and reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6779–6789, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.452. URL https://aclanthology.org/2023.findings-emnlp.452.

[20] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2024. URL https://arxiv.org/abs/2404.16130.

[21] Mohamed Elhoseiny and Ahmed M. Elgammal. Text to multi-level mindmaps: A new way for interactive visualization and summarization of natural language text. *CoRR*, abs/1408.1031, 2014. URL http://arxiv.org/abs/1408.1031.

[22] Theodore Dalamagas, Tryfon Farmakakis, Manolis Maragkakis, and Artemis G. Hatzigeorgiou. Freepub: Collecting and organizing scientific material using mindmaps. *ArXiv*, abs/1012.1623, 2010. URL https://api.semanticscholar.org/CorpusID:12397418.

[23] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL http://arxiv.org/abs/1707.06347.

[24] Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. HoVer: A dataset for many-hop fact extraction and claim verification. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.309. URL https://aclanthology.org/2020.findings-emnlp.309.

[25] Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. Feverous: Fact extraction and verification over unstructured and structured information. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/68d30a9594728bc39aa24be94b319d21-Paper-round1.pdf.

[26] David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-open: Towards open-domain scientific claim verification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.347. URL https://aclanthology.org/2022.findings-emnlp.347.

[27] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580.

[28] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies, 2021. URL http://arxiv.org/abs/2101.02235.

[29] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31.

[30] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

[31] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *CoRR*, abs/1710.06481, 2017. URL http://arxiv.org/abs/1710.06481.

[32] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL http://arxiv.org/abs/2203.02155.

[33] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1387. URL https://aclanthology.org/D19-1387.

 https://doi.org/10.52202/079017-4137

A Prompt for Extraction

The specific prompt of extraction is shown in Table 6. The purpose of the prompt is to obtain logical relationships and multi-hop connections from the context. In the prompt, the content within the brackets should be replaced with specific example content.

Table 6: Specific prompt for obtaining logical relationships. During execution, "[EVIDENCE]" needs to be replaced with the specific document, and "[CLAIM]" is replaced with the specific question.

Prompt Content for Logical Relationship Extraction
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]

To clearly demonstrate the content of the LLM after performing the extraction, we give an example in the MuSiQue dataset.

Document: 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. The original music score is by Miklos Rozsa. The film stars Marlon Brando as Mark Antony, James Mason as Brutus, John Gielgud as Cassius, Louis Calhern as Julius Caesar, Edmond O'Brien as Casca, Greer Garson as Calpurnia, and Deborah Kerr as Portia. Houseman was born on September 22, 1902, in Bucharest, Romania, the son of May (Davies) and Georges Haussmann, who ran a grain business. His mother was British, from a Christian family of Welsh and Irish descent. His father was an Alsatian-born Jew. 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.

JSON format after extraction:


```
{
  "Julius Caesar": {
    "Production Company": "Metro-Goldwyn-Mayer",
    "Adaptation": "Play by Shakespeare",
    "Director": "Joseph L. Mankiewicz",
    "Screenplay": {
        "Writer": "Joseph L. Mankiewicz",
        "Credit": "Uncredited"
        },
    "Producer": {
        "John Houseman": {
            "Birth": {
                "Date": "September 22, 1902",
                "Place": "Bucharest, Romania"
            },
            "Parents": {
                "Mother": {
                    "Name": "May Davies",
                    "Nationality": "British",
                    "Religion": "Christian",
                    "Descent": "Welsh and Irish"
                },
                "Father": {
                    "Name": "Georges Haussmann",
                    "Occupation": "Grain Business",
                    "Religion": "Jew",
```

```
                    "Birthplace": "Alsace"
                }
            },
            "Education": "Clifton College, England",
            "Citizenship": {
                "Original": "British",
                "Naturalized": "United States, 1943"
            },
            "Occupation": {
                "Original": "Grain Trade, London",
                "Stage Name": "John Houseman"
            },
            "Emigration": "United States, 1925"
        }
    },
    "Music Score": {
        "Composer": "Miklos Rozsa"
    },
    "Cast": {
        "Marlon Brando": "Mark Antony",
        "James Mason": "Brutus",
        "John Gielgud": "Cassius",
        "Louis Calhern": "Julius Caesar",
        "Edmond O'Brien": "Casca",
        "Greer Garson": "Calpurnia",
        "Deborah Kerr": "Portia"
    }
  }
}
```


B Input Format For Pruning

The extracted context g includes logical relationships and corresponding attribute values. First, we iterate through all logical relationships and attribute values, and following the previous method [33], we concatenate them using [SEP] [CLS], combining them with the question to input into the pruning model. Then, we use the representation of the [CLS] token to represent the logical relationships for subsequent operations.

We use an example in Figure 2 to demonstrate the input format of the extracted context g in the pruning model. The question is: Where did the producer of Julius Caesar study or work? The the extracted context g is:

```
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
```

We only traverse logical relationships of the first-level progressive type. Therefore, the logical relationships contained in extracted context g of example include 1) Production Company: Metro-Goldwyn-Mayer, 2) Director: Joseph L. Mankiewicz, 3) Producer: Name: John Houseman, 4) Producer: Education: Clifton College,England, 5) Producer: Occupation: Grain Trade, London, 6) Adaptation: Play by Shakespeare. Then, we concatenate them using [SEP][CLS].

The input format after concatenation is: [CLS] Production Company: Metro-Goldwyn-Mayer [SEP] [CLS] Director: Joseph L. Mankiewicz [SEP] [CLS] [CLS] Producer: Name: John Houseman [SEP]

 https://doi.org/10.52202/079017-4137

[SEP] [CLS] Producer: Education: Clifton College [SEP] [CLS] Producer: Occupation: Grain Trade, London [SEP] [CLS] Adaptation: Play by Shakespeare [SEP] [CLS] Where did the producer of Julius Caesar study or work? [SEP].

C Prompt for Multi-hop Reason

During the reasoning stage using large language models, to accommodate more context-aware reasoning tasks while ensuring comparability of results, we designed a universal prompt template. The prompt template consists of three components: original context, e.g., documents or paragraphs, reorganized information, and a question. The prompt template is as follows:

```
Documents:
   [Re-Organized TEXT]
   [TEXT] [Optional]
Question:
   [QUESTION]
please answer the question based on the documents.
Answer:
```

The specific prompts for Standard, chain-of-thought (CoT), our InfoRE, and InfoRE + CoT to reason are shown in Table 7. In the prompt, the content within the brackets [] should be replaced with specific example content.

Table 7: Specific prompts and answer format instructions of Standard, CoT, InforRE, and InfoRE + CoT in our paper. During execution, "[EVIDENCE]" needs to be replaced with the specific document, "[CLAIM]" is replaced with the specific question, and "[MindMap]" is replaced with the specific MindMap.

Methods	**Prompt Content**
Standard	Documents: [EVIDENCE] Question: [CLAIM]? Please answer the question based on Documents. Your final answer should be enclosed in XML tag <answer></answer>, like this: <answer>{*final_answer*}</answer>, at the end of your response. Answer:
CoT	Documents: [EVIDENCE] Question: [CLAIM]? Please answer the question based on Documents. Your final answer should be enclosed in XML tag <answer></answer>, like this: <answer>{*final_answer*}</answer>, at the end of your response. Let's think step by step. Answer:
InfoRE	Documents: [MindMap] Question: [CLAIM]? Please answer the question based on Documents. Your final answer should be enclosed in XML tag <answer></answer>, like this: <answer>{*final_answer*}</answer>, at the end of your response. Answer:
InfoRE + CoT	Documents: [MindMap] Question: [CLAIM]? Please answer the question based on Documents. Your final answer should be enclosed in XML tag <answer></answer>, like this: <answer>{*final_answer*}</answer>, at the end of your response. Let's think step by step. Answer:

D Detailed Dataset Information

In our experiments, due to the resource limitations of large language models, we sample a portion from each dataset, following previous methods [6, 16].

Table 8: Details statics information of evaluation datasets we used in the paper. Pairs denote the number of examples. 2WMHopQA is the short name of 2WikiMultiHopQA.

Tasks	**Dataset**	**Pairs**	
		train	**test**
Claim Verification	FEVEROUS	2000	2959
	HOVER	2000	4000
	SCIFACT	200	212
Question Answering	2WikiMultiHopQA	2000	500
	StrategyQA	1000	229
	MusiQue	2000	2417
Reading Comprehension	HotpotQA	2000	500
	WIKIHOP	2000	500

Claim Verification The task involves assessing a given claim against a set of evidence documents to determine whether they support or refute the claim [12]. We consider HOVER [24] and FEVEROUS [25], which comprise complex claims that necessitate multi-hop reasoning for verification. Besides, we also take into account the SCIFACT [26] dataset, notable for its inclusion of scientific claims.

Question Answering For this task, we consider the following datasets: 2WikiMultiHopQA [27], StrategyQA [28], MuSiQue [29], and HotpotQA [30]. To answer the questions in these datasets requires not only multi-hop reasoning but also cross-document analysis.

Reading Comprehension Machine reading comprehension task requires a model to process documents and select an answer from the provided candidates to a question about the content [14]. We primarily consider WIKIHOP [31] in the task, which necessitates multi-hop reasoning to derive the final answer. Additionally, HotpotQA dataset [30] is also frequently considered as part of the question answering domain.

E Limitations

We propose an information re-organization approach to improve the reasoning of large language models, which performs well on some context-aware reasoning tasks but still has some limitations. Firstly, the structures of information re-organization are limited. Generally, there are multiple structures for information reorganization, such as tables, timelines, etc. Next, we will extend the re-organization structure to include more types. Secondly, the re-organization process relies on large language models. If we can implement this re-organization using smaller language models, our method will become more generalizable. This is another direction we need to focus on in the future.

https://doi.org/10.52202/079017-4137

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract, Introduction

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations in Appendix

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-4137

Answer: [Yes]

Justification: Supplementary material

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because it would be too computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experimental

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Supplementary material

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Conclusion

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

https://doi.org/10.52202/079017-4137

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Introduction, Related Work, Methodology, Experimental, Results and Analysis.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not involve the release of new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

 https://doi.org/10.52202/079017-4137