PREMISES REORDERING IN FORWARD CHAINING IMPROVES LLM SYMBOLIC REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have shown outstanding performance on diverse natural language processing tasks, but they still struggle with complex logical reasoning, limiting its real-world applicability. While previous neuro-symbolic approaches for improving LLMs' performance on logical question-answering (QA) primarily focus on either translation quality or reasoning process, they largely overlook LLMs' performance significantly sensitive to the **orders** of relevant information (also known as **premises** in logical QA tasks) in the input contexts. Motivated by such observations, we propose a method to first reorder the logical QA premises to align with the premises orders in forward chaining proof to improve LLM logical reasoning. We then use the LLM to translate both premises and the question to the symbolic language, and perform symbolic reasoning via an external logic solver using the translated symbolic language. In this way, both the translation and reasoning accuracy are enhanced, due to the forward chaining premises benefits (i) sequentially formulating the objects and (ii) performing symbolic reasoning during the premises searching in solver. Empirical experiments across three benchmarks demonstrate that our premises reordering method stably outperforms neuro-symbolic baselines, including both symbolic solver-based and prompt-based methods.

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable performance on various natural language processing tasks, but still struggle in complex logical reasoning, which significantly limits their practical applicability in real-world scenarios (Cheng et al., 2025). To evaluate LLMs' logical reasoning capabilities, logical question answering (QA) task required LLMs to decide whether a statement can be logically deduced from the given information (which is known as the premise). The LLM is expected to distinguish whether the specific statement is *true* (it can be inferred from the premises), *false* (it can be proved to contradict the premises), or *unknown* (when the given premises are insufficient to infer or refute it) with the premises.

There are two neuro-symbolic paradigms in existing methods to improve the logical reasoning capabilities of LLMs: symbolic solver-based methods (Olausson et al., 2023; Ye et al., 2023; Ryu et al., 2025) and prompt based methods (Liu et al., 2025; Ozeki et al., 2024; Xu et al., 2024b; 2025). Symbolic solver-based approaches firstly translate natural language (NL) queries into symbolic language (SL) by LLMs, and then employ an external logical symbolic solver to process the logical reasoning in these symbolic representations. Alternatively, prompt-based strategies leverage LLMs to perform the translation, reasoning and verification processes via designed prompts.

However, previous methodologies primarily focus on improving translation quality and ensuring LLMs apply appropriate reasoning rules, largely overlooking **the order of premises** can also significantly impact the performance of LLMs in logical QA tasks. Though in logical and mathematical reasoning theory, changing the order of premises alone does not affect the validity of the conclusion, it has been experimentally proven that LLMs achieve optimal performance when the premise order in the context aligns with the intermediate steps of the ground-truth forward reasoning chain Chen et al. (2024); He et al. (2025). For example, as shown in Figure 1, for the same question with two different premise orders, the model answers correctly when the premise order matches the ground-truth proof (right purple box), but answers wrongly under the original disordered premise sequence

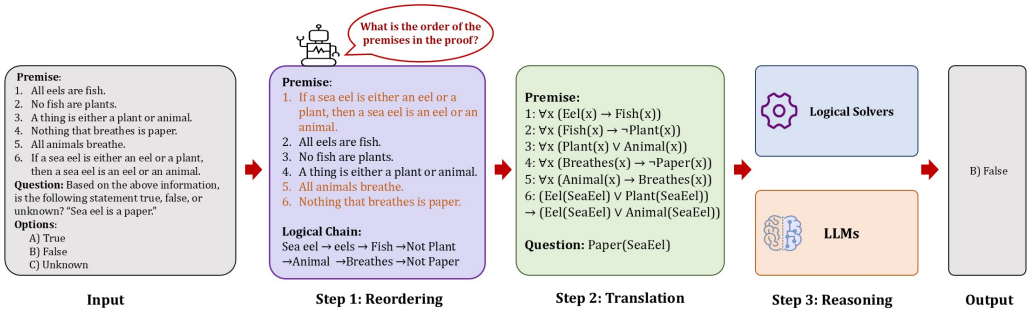

Figure 1: The overall framework of our proposed method. We first reorder the input premises to align with the ground-truth forward-chaining proof path (Step 1). Subsequently, the structured context undergoes the translation and reasoning process (Step 2 & 3), compatible with both symbolic solver-based and prompt-based methods to derive the final answer.

(left gray box). In general, presenting premises in the order that mirrors the ground-truth deduction proof significantly reduces hallucination, while the performance drops by over 30% Chen et al. (2024) if LLMs are forced to reason over disjointed or shuffled premises.

Motivated by this observation, we propose a novel framework that reorder the given premises based on the ground-truth deduction proof with forward chaining for each QA problem before the translation and reasoning process. Specifically, we leverage the LLM to restructure the input premises (contexts) into the order that mirrors the proof path where dependencies are presented sequentially (e.g., placing "If A then B" prior to "If B then C"). Following this reordering stage, the translation and reasoning processes are then executed. By organizing premises to a forward-chaining flow, we enhance both translation and reasoning accuracy, since the reordered premises is beneficial to not only formalize the entities in NL sequentially, but also perform symbolic reasoning step-by-step. Our experiments demonstrate that our premises reordering method consistently improves performance of LLMs across various baseline methods. Our main contributions are summarized as follows:

- We propose the first method to explicitly reorder the premises based on the ground-truth proof before the translation and reasoning process, to improve LLMs' performance in logical QA.
- We empirically validate our method through diverse experiments, showing that our premises reordering method consistently outperforms all the baseline methods.
- We demonstrate the generalization of our approach by integrating it into both symbolic solver-based and prompt-based paradigms with better performance.

2 Related Work

2.1 Logical Reasoning in LLMs

Current methodologies for enhancing LLM logical reasoning generally fall into three categories. Prompt-based methods employ in-context strategies like Chain-of-Thought (CoT) to guide step-by-step deduction in symbolic representations Xu et al. (2024b; 2025), while fine-tuning approaches train LLMs on synthetic datasets to internalize logical reasoning patterns Morishita et al. (2024); Wan et al. (2024). Alternatively, solver-based methods use LLMs to perform NL-to-SL translation for external solvers Ye et al. (2023); Ryu et al. (2025) to perform reasoning. Unlike existing frameworks that manage to improve the reasoning engine or translation quality, our work is the first to introduce premise reordering before translation and reasoning process.

Table 1: Performance comparison of logical reasoning accuracy between Deepseek-v3 and GPT-4 across three benchmarks (ProntoQA, ProofWriter, and LogiDeduction). We compare five baseline methods under two settings: original input ordering (w/o reorder) and our proposed forward-chaining reordering (w reorder).

		Deepseek-v3			**GPT-4**		
		ProntoQA	ProofWriter	LogicDeduction	ProntoQA	ProofWriter	LogicDeduction
CoT	w/o reorder	98.20%	90.17%	92.00%	91.20%	78.67%	88.67%
	w/ reorder	**98.80%**	**91.33%**	**94.00%**	**92.60%**	**80.83%**	**91.67%**
LogicLM	w/o reorder	83.20%	80.50%	**93.33%**	**93.40%**	79.17%	87.00%
	w/ reorder	**84.00%**	**80.67%**	**93.33%**	**93.40%**	**79.33%**	**87.33%**
SymbCoT	w/o reorder	98.00%	85.83%	94.00%	96.00%	82.33%	86.33%
	w/ reorder	**98.40%**	**86.33%**	**94.33%**	**96.20%**	**83.67%**	**87.00%**
CR	w/o reorder	95.40%	80.33%	83.67%	93.20%	71.67%	80.33%
	w/ reorder	**95.80%**	**81.83%**	**84.00%**	**93.60%**	**72.17%**	**81.00%**
DetermLR	w/o reorder	96.80%	**82.17%**	88.33%	**97.80%**	77.33%	85.00%
	w/ reorder	**97.20%**	**82.17%**	**88.67%**	97.60%	**78.00%**	**85.33%**

2.2 Order Sensitivity in LLMs

Although classical logic is theoretically invariant to premise order, LLMs exhibit significant sensitivity to the input sequence. Empirical studies highlight that the position of relevant information drastically affects performance, often citing the "lost-in-the-middle" effect Liu et al. (2024). In logical reasoning specifically, LLMs achieve optimal accuracy when premises align with the order of the ground-truth reasoning chain, whereas randomized inputs cause substantial degradation Chen et al. (2024); He et al. (2025). While previous works primarily analyze order sensitivity as a model vulnerability, we propose a method to reorder the input to achieve better performance in LLMs' logical reasoning.

3 Proposed Method

As illustrated in Figure 1, our framework enhances logical reasoning of LLMs by restructuring the input context before reasoning. The pipeline consists of two main stages: explicitly reordering premises to align with the ground-truth forward-chaining proof path, followed by the translation and reasoning process compatible with both solver-based and prompt-based paradigms.

3.1 Premise Reordering based on Forward Chaining

To mitigate the performance degradation caused by disordered or random contexts, we introduce a adaptive method that reorders the input premises to mirror the ground-truth deduction proof. Specifically, we prompt an LLM to analyze the logical dependencies within the context and reorder the premises based on forward chaining proof, where a premise $A \rightarrow B$ is positioned strictly before $B \rightarrow C$ to infer the conclusion $A \rightarrow C$. It can be refined by generating a coherent "logical chain"—such as tracing from the entity in the question to its property (e.g., *Sea eel* $\rightarrow$ *Fish* $\rightarrow \ldots \rightarrow$ *Not Paper*). In this way, we effectively reduce the workload and complexity stemmed from information retrieval. This alignment ensures that the subsequent translation and reasoning engine process structured premises rather than a set of disjointed or random ordered facts, thereby minimizing hallucinations and logical disconnects.

3.2 Translation and Reasoning

Following the reordering stage, the structured premises undergo a translation and reasoning process that generalizes across different neuro-symbolic approaches. First, the reordered NL context is translated into formal SL, such as First-Order Logic, by an LLM. Subsequently, the reasoning step is executed to determine the truth value of the conclusion (*True*, *False*, or *Unknown*). This stage can either offload the symbolic representations to an external logical solver (symbolic solver-based method) for rigorous proofs or leverage the LLM itself (prompt-based method) to perform

reasoning over the reordered context. By decoupling the ordering optimization from the reasoning process, our method consistently enhances performance regardless of the downstream reasoning engine employed.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

To evaluate the effectiveness of our proposed reordering method, we conducted experiments across three benchmark datasets of logical reasoning: **ProntoQA** Saparov & He (2023), **ProofWriter** Tafjord et al. (2021), and **LogicDeduction** Srivastava et al. (2023). These datasets encompass varying degrees of complexity, ranging from simple deductive chains to complex multi-hop reasoning scenarios involving negation and quantification.

We utilized two LLMs as backbones: **Deepseek-v3** and **GPT-4**. We compared our approach against five established baselines, which are categorized into two distinct paradigms based on their reasoning execution:

- **Solver-based Method: LogicLM** Pan et al. (2023). This framework operates by translating NL into SL (e.g., Prolog or FOL) and performs the inference process by the corresponding solvers.
- **Prompt-based Methods:** This category includes **Chain-of-Thought (CoT)** Wei et al. (2022), **Symbolic Chain-of-Thought (SymbCoT)** Xu et al. (2024a), **Cumulative Reasoning (CR)** Zhang et al. (2023), and **DetermLR** Sun et al. (2024). These methods rely entirely on the LLMs to perform NL-to-SL translation and step-by-step logical deduction.

For all baselines, we conduct experiments under two conditions: standard input ordering (*w/o reorder*) and our reordering (*w reorder*) based on forward chaining proof.

4.2 RESULTS ANALYSIS

Table 1 presents the comparative performance of five baselines across the three datasets. The results demonstrate that our reordering method enhances performance across all models and baselines. Note that the improvement differs between the two baseline paradigms.

Impact on Prompt-based Methods. Methods relying on LLMs reasoning (e.g., CoT, SymbCoT, CR) exhibit a more pronounced sensitivity to the premise ordering. As observed in Table 1, applying reordering to CoT on the ProofWriter dataset achieves notable gains (e.g., GPT-4 performance improves from 78.67% to 80.83%). By aligning the input sequence with the ground-truth reasoning chain, our method facilitates the generation of reasoning paths, reducing hallucination and workload long-distance retrieval during intermediate deduction steps.

Impact on Solver-based Methods. Solver-based approach (LogicLM) shows a relatively smaller performance variance between the *w/o reorder* and *w reorder* settings compared to the prompting methods. This is resulted from that the reasoning engine of LogicLM is an external solver (e.g., Pyke or Z3), which is theoretically invariant to the order of logical formulas (i.e., $\{A, B\} \vdash C$ is equivalently complex to $\{B, A\} \vdash C$). Therefore, the reordering does not affect the reasoning stage much. Instead, the observed gains stem from the improved translation qualities. A logically coherent premise order benefits the LLMs' performance during the translation stage (NL $\rightarrow$ SL), ensuring higher consistency in predicate definition and variable binding, thereby generating syntactically correct and semantically accurate symbolic expressions for the solver.

5 CASE STUDY

To intuitively demonstrate the critical impact of premise ordering on logical reasoning in LLMs, we conduct a detailed case study on a complex instance from the ProofWriter dataset (ID: RelNeg-OWA-D5-922_Q7). This problem requires multi-hop deductive reasoning involving negation and quantifiers to determine the truth value of the statement: *"The mouse eats the mouse."*

The ground truth is *True*. The validity of this conclusion relies on the folowing logical reasoning chain:

1. Fact: `Sees(cow, squirrel)`
2. Rule 1: `Sees(cow, squirrel)` $\rightarrow$ `Eats(cow, mouse)`
3. Rule 2: `Eats(x, mouse)` $\rightarrow$ `Blue(x)` (can infer `Blue(cow)`)
4. Rule 3: `Needs(x, cow)` $\wedge$ `Blue(cow)` $\rightarrow$ `Eats(x, mouse)`
5. Fact: `Needs(mouse, cow)`
6. Conclusion: Combine Fact 5 and inferred `Blue(cow)` to derive `Eats(mouse, mouse)`.

We compare the reasoning process generated under the original disordered context versus our re-ordered context based on the forward chain proof.

Failure with Original Ordering. In the original setting, the premises are presented in a random, disjointed sequence. For example, the crucial rule defining the "Blue" attribute appears at the very end of the context (Premise 19), far removed from the initial facts. As shown in the trace below, the model exhibits a "reasoning cut-off" because it fails to bridge the gap to the intermediate variable `Blue(cow)`:

```
# Step: Attempting to use Premise 18
# Rule: Needs(x, cow) & Blue(cow) -> Eats(x, mouse)

Reasoning Trace:
 Premise 18 states: Forall x (Needs(x, cow) & Blue(cow)) -> Eats(x,
    ↪ mouse).
But Blue(cow) is not given in the premises...
We have Blue(tiger), but not Blue(cow).
So we cannot use this rule to derive Eats(mouse, mouse).

# Result: Deduction Fails
```

Due to the disordered presentation, the model fails to trigger the backward chain to deduce `Blue(cow)` from `Eats(cow, mouse)`. Consequently, it incorrectly concludes that the statement is unprovable, resulting in the wrong answer *Unknown*.

Success with Forward Chain Reordering. Our method reorganizes the premises to align with the forward-chaining flow of the ground-truth proof. By placing dependencies closer to their triggers (e.g., positioning rules about "eating the mouse" and "being blue" in a coherent sequence), the reordering reduces the search workload for the model. The reasoning trace below demonstrates a successful derivation:

```
# Step 1: Inferring Intermediate Facts
 From premise 15: Sees(cow, squirrel) is true.
  -> Therefore, Eats(cow, mouse) is true.

# Step 2: Chaining to Attribute 'Blue'
 From premise 1: Eats(x, mouse) -> Blue(x).
  -> Applying to cow: Blue(cow) is true.

# Step 3: Final Deduction
 Apply Premise 19 to x = mouse:
Needs(mouse, cow) & Blue(cow) -> Eats(mouse, mouse).
  -> We have Needs(mouse, cow) (Premise 8).
  -> We have derived Blue(cow) (Step 2).
  -> Thus, Eats(mouse, mouse) is true.
```

This comparison highlights that the difficulty in LLM logical reasoning often lies not in the inability to understand individual rules, but in the retrieval failure of necessary premises during multi-step

deduction. Our reordering mechanism mitigates this by reordering the premises to guide the model along the ground-truth proof path.

6 CONCLUSION

In this work, we introduced a novel framework that explicitly reorders input premises to align with the ground-truth forward-chaining proof prior to the translation and reasoning processes. By restructuring disordered or random contexts into coherent logical chains, our method effectively mitigates the retrieval workload and reduces the "lost-in-the-middle" effect during multi-hop logical reasoning. Our extensive experiments demonstrate that aligning premises with the forward reasoning path consistently enhances performance for both symbolic solver-based and prompt-based paradigms. Notably, our analysis further reveals that while reordering directly guides the reasoning path in prompt-based methods, it also benefits solver-based approaches primarily by improving the NL-to-SL translation quality.

REFERENCES

Xinyun Chen, Ryan A Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 6596–6620, 2024.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. Empowering llms with logical reasoning: A comprehensive survey. *International Joint Conference on Artificial Intelligence, Survey Track*, 2025.

Qianxi He, Qianyu He, Jiaqing Liang, Weikang Zhou, Zeye Sun, Fei Yu, and Yanghua Xiao. Order doesn't matter, but reasoning does: Training LLMs with order-centric augmentation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27166–27180, 2025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, pp. 157–173, 2024.

Tongxuan Liu, Wenjiang Xu, Weizhe Huang, Yuting Zeng, Jiaxing Wang, Xingyu Wang, Hailong Yang, and Jing Li. Logic-of-thought: Injecting logic into contexts for full reasoning in large language models. In *Proceedings of Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. *Advances in Neural Information Processing Systems*, 2024.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2023.

Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. Exploring reasoning biases in large language models through syllogism: Insights from the NeuBAROCO dataset. In *Findings of ACL*, 2024.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Hyun Ryu, Gyeongman Kim, Hyemin S Lee, and Eunho Yang. Divide and translate: Compositional first-order logic translation and verification for complex logical reasoning. In *The International Conference on Learning Representations*, 2025.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Hongda Sun, Weikai Xu, Wei Liu, Jian Luan, Bin Wang, Shuo Shang, Ji-Rong Wen, and Rui Yan. DetermLR: Augmenting LLM-based logical reasoning from indeterminacy to determinacy. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.531.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of ACL*, 2021.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael R Lyu. Logicasker: Evaluating and improving the logical reasoning ability of large language models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 2022.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2024a.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2024b.

Jundong Xu, Hao Fei, Meng Luo, Qian Liu, Liangming Pan, William Yang Wang, Preslav Nakov, Mong-Li Lee, and Wynne Hsu. Aristotle: Mastering logical reasoning with a logic-complete decompose-search-resolve framework. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2025.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. Satlm: Satisfiability-aided language models using declarative prompting. *Advances in Neural Information Processing Systems*, 2023.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. Cumulative reasoning with large language models. *arXiv preprint arXiv:2308.04371*, 2023.