Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Doctoral Consortium Track

Enhancing the Logical Reasoning Abilities of Large Language Models

Fengxiang Cheng
Institute for Logic, Language and Computation (ILLC), University of Amsterdam
f.cheng@uva.nl

Abstract

Large language models (LLMs) have demonstrated impressive progress in various natural language progress tasks. However, it has been observed that LLMs still struggle with complex causal and logical reasoning. To facilitate this research direction, we first proposed a training method to distinguish causal relationships from spurious correlations in sentiment classification tasks. Then we conducted a comprehensive survey categorizing existing approaches, firstly identifying the main challenges of complex logical question-answering tasks and logical inconsistency across different questions. Our ongoing projects mainly focus on two points: (1) incorporating modal and epistemic logic to evaluate and enhance LLMs' reasoning ability to handle more complex and diverse reasoning tasks, and (2) phased training LLMs with curriculum learning to improve their logical reasoning performance.

1 Introduction and Motivation

Large language models (LLMs) have achieved remarkable advances across a diverse range of natural language tasks. However, it has been revealed that there remain substantial challenges with respect to the complex causal and logical reasoning capabilities of LLMs. This can stem from the fact that learning languages and knowledge through tasks such as next word prediction or masked language modeling fail to ensure the accuracy of causal and logical reasoning, and the limitation that the pre-training corpus of LLMs lacks high-quality logical reasoning examples [Morishita *et al.*, 2024]. These challenges severely restrict the applicability of LLMs in both complex real-world scenarios and high-stakes situations.

Numerous efforts have been made to enhance the reasoning capabilities of LLMs such as chain-of-thought (CoT) and more recently, reasoning LLMs like DeepSeek-R1 have emerged. However, studies explicitly focused on enhancing logical reasoning capabilities remains rather limited, and the field of improving LLM logical reasoning abilities is still in its early stages, facing several crucial challenges and corresponding research questions (RQs): **RQ1: Can we comprehensively summarize and categorize difficulties and methodologies about LLM's logical reasoning**? Next, most of the existing methods are limited to propositional and first-order logic, leading to **RQ2: Can we extend existing methods that improves LLMs logical reasoning to solve modal logic, epistemic logic or more complex logical reasoning problems?** Lastly, we find that fine-tuning LLMs directly with logical corpora leads to sub-optimal performance, as many complex logical inference rules are difficult to be understood directly by LLMs. To address this issue, we aim to answer **RQ3: Can we gradually fine-tune the LLMs from simple to complex logic samples to further enhance their logical reasoning abilities?** Therefore, our research will focus on the above three RQs to enhance the logical reasoning capabilities of LLMs.

2 Previous Contributions

Existing inference of LLMs primarily relies on retrieving information from their training corpus, following correlation-based rules. However, due to the lack of explicit data on complex reasoning tasks—particularly in causal and logical reasoning, language models often experience hallucinations resulted from spurious correlations. For instance, in the IMDB movie reviews dataset, the word "and" exhibits a stronger correlation with positive sentiment than the word "excellent", despite "and" clearly lacking intrinsic sentiment polarity. To address this issue, in our previous work, we introduced probability of necessity (PN) and probability of sufficiency (PS) to answer the counterfactual question "If a sentence has a certain sentiment in the presence/absence of a word, would the sentiment change when that word is absent/present?" [Cheng *et al.*, 2025a]. Our approach trains language models to distinguish causal relationships from spurious correlations in sentiment classification tasks, particularly in cases where multiple causal words appear within the same sentence.

To answer RQ1, we conducted a comprehensive and systematic survey of the logical reasoning of LLMs [Cheng *et al.*, 2025b]. Our paper summarizes and categorizes the main challenges into two aspects: (1) LLMs often fail to generate the correct answer *within* a complex **logical question answering** task that demand sophisticated deductive, inductive or abductive reasoning based on a set of premises or constraints. For instance, the LLaMA-13B model attains an accuracy of 33.63% when prompted with 8-shot on the logical questions dataset FOLIO and it is only marginally superior to the random guessing among true, false, and unknown options.

This severely limits the application of LLMs in complex real-world scenarios tasks. (2) LLMs are also inclined to produce responses contradicting themselves, the knowledge base, or logical rules *across* different questions, which is considered as a violation of **logical consistency**. For example, a state-of-the-art Macaw question-answering LLM answers *Yes* to both questions *Is a magpie a bird?* and *Does a bird have wings?* but answers *No* to *Does a magpie have wings?*. Consequently, these conflicting outputs raise concerns about the reliability and trustworthiness of LLMs, restricting their practical implementation, particularly in high-stakes situations.

To promote this research direction, we thoroughly investigate the latest methods and put forward detailed taxonomies of them. For logical question answering, the corresponding methods are categorized into solver-based, prompt-based, and fine-tuning methods. Specifically, solver-based methods first convert problems in natural language into symbolic language expressions, and then solve them using external logical solvers. Prompt-based methods includes two branches: (1) explicitly modeling the logical chains of question-answering task, and (2) prompting the LLM to translate natural language into symbolic language, make inference and verify the results [Xu *et al.*, 2024]. Fine-tuning methods train LLMs with augmented data containing deductive proofs and natural language examples that explicitly consist of the logical reasoning process. Regarding logical consistency, we first distinguish the most common categories of logical consistency, including negation, implication, transitivity, factuality consistency, and their composites. For each type, we discuss the corresponding advanced solutions. Lastly, we summarize commonly used benchmark datasets and evaluation metrics for these two aspects.

To the best of our knowledge, this is the first work to comprehensively investigate the most cutting-edge research on enhancing the logical reasoning capabilities of LLMs, including correctly answering complex logical questions and improving various logical consistency across their answers to different questions.

3 Directions for Future Work

In this section, we extend our interest of LLMs reasoning ability to modal and epistemic logical reasoning (RQ2) and introduce curriculum learning to fine-tune the LLMs to improve their logical reasoning capabilities (RQ3).

To answer RQ2, we focus on modal logic and epistemic logic, to evaluate and improve the ability of LLMs to handle more complex and diverse reasoning tasks. Modal logic extends propositional logic by introducing operators "must" ($\Box$) and "may" ($\Diamond$) to express certainty and possibility, respectively. For example, *Mary **might** ($\Diamond$) not ($\neg$) get the perfect score (p)* implies *It's not ($\neg$) the case that Mary **must** ($\Box$) get the perfect score (p)*, which can be formally formulated as $\Diamond\neg p \vDash \neg\Box p$. Similarly, epistemic logic formalizes knowledge and beliefs of agents, enabling to model the dynamic cognitive states and reasoning processes of multi-agents in complex interactive scenarios such as games, especially when the information updates. It provides a framework for training LLMs to solve complex cognitive puzzles involving belief revision and knowledge updates during interactions of multi-agents. To this end, we are developing benchmarks to systematically assess the capabilities of LLMs in modal and epistemic logic reasoning, based on which we further propose new methods to improve their performance in solving challenging cognitive puzzles.

To answer RQ3, we leverage curriculum learning methods to train LLMs gradually with logical reasoning samples from easy to hard to enhance their logical reasoning capabilities. Curriculum learning is a training paradigm inspired by human learning processes, where the model first learns the simpler and then the more complex tasks. This approach has been shown to improve models' performance across various machine learning domains, including natural language processing, computer vision, and reinforcement learning. Recently, it has been applied to LLMs' instruction fine-tuning, demonstrating significant performance gains. Motivated by such findings, we propose to gradually fine-tune LLMs in logical reasoning via adaptively selecting samples from easy to hard. The superiority of our approach will be evaluated on commonly used logical reasoning benchmarks.

4 Conclusion

Overall, with the observation that LLMs are struggling with complex causal and logical reasoning, we first proposed a method for training language models to distinguish causal relationships from spurious correlations in sentiment classification tasks. Then we conducted a comprehensive survey of the latest studies focusing on logical reasoning abilities of LLMs, identifying two main challenges and categorized existing approaches in this field. For future work to address the challenges, we will focus on two primary directions: (1) advancing modal and epistemic logic reasoning by establishing benchmarks and proposing new approaches, and (2) incorporating curriculum learning to gradually fine-tune LLMs to improve their logical reasoning capabilities. Our research aims to make contributions to improving the logical reasoning abilities of LLMs, including achieving better performance within complex logical question-answering tasks and ensuring logical consistency across LLMs' outputs.

References

[Cheng *et al.*, 2025a] Fengxiang Cheng, Haoxuan Li, Alina Leidinger, and Robert Van Rooij. Revealing the limitations of exploiting causal effects to resolve linguistic spurious correlations. In *AAAI 2025 Workshop on AICT*, 2025.

[Cheng *et al.*, 2025b] Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. Empowering LLMs with logical reasoning: A comprehensive survey. In *IJCAI*, 2025.

[Morishita *et al.*, 2024] Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Enhancing reasoning capabilities of LLMs via principled synthetic logic corpus. In *NeurIPS*, 2024.

[Xu *et al.*, 2024] Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In *ACL*, 2024.