Fine-Tuning Sample Order Matters in Propositional Logical Question-Answering (Student Abstract)

Fengxiang Cheng[1], **Chuan Zhou**[2], **Fenrong Liu**[3,1*], **Robert van Rooij**[1*]
[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]School of Mathematics and Statistics, The University of Melbourne
[3]Tsinghua-UvA JRC for Logic, Tsinghua University
{f.cheng, r.a.m.vanrooij}@uva.nl, chuan.zhou@student.unimelb.edu.au, fenrong@tsinghua.edu.cn

Abstract

Large language models (LLMs) have achieved impressive progress in natural language processing tasks but still struggle with complex logical reasoning. We observe that in propositional logic question-answering (QA), LLMs' performance varies with the order of training samples during fine-tuning. Motivated by this, we propose a data-driven approach to automatically determine the fine-tuning sample order, enhancing the logical QA performance of LLMs. Specifically, we first quantify the logical reasoning complexity of propositional reasoning samples and then stratify the training data into several subsets of ascending complexity. Subsequently, we fine-tune the LLMs on these subsets, progressing from low to high reasoning complexity. Experimental results demonstrate that our approach outperforms single-stage fine-tuning baselines across diverse reasoning benchmarks.

1 Introduction

Large language models (LLMs) have exhibited remarkable performance in a large number of natural language processing (NLP) tasks. However, research indicates that LLMs still struggle with complex scenarios that demand logical reasoning capabilities (Cheng et al. 2025). Previous methods for improving logical question answering (QA) of LLMs can be broadly divided into three categories: external solver-based methods (Olausson et al. 2023; Pan et al. 2023; Ye et al. 2023), prompt-based methods (Xu et al. 2024; Liu et al. 2025), and fine-tuning methods (Feng et al. 2024; Morishita et al. 2024; Wan et al. 2024).

Despite these advancements, each pipeline faces distinct limitations. Solver-based approaches are highly susceptible to translation errors (Ryu et al. 2025; Wang et al. 2025), while prompt-based techniques remain constrained by the model's intrinsic reasoning abilities (Xu et al. 2025; Cheng 2026). Furthermore, we observe that direct fine-tuning often fails to resolve complex logical reasoning challenges effectively, which can be stemmed from the inability of LLMs to comprehend logical rules and struggle to simultaneously process multiple premises (Cheng 2025; Fu et al. 2025).

To address these limitations, we propose a **reasoning complexity-based fine-tuning approach**, by noting that fine-tuning sample order matters in LLMs' performance in propositional logical QA. We stratify training dataset of propositional logic in ascending order of reasoning complexity which is quantified by an LLM. Then we fine-tune LLMs on these subsets of training data from low to high reasoning complexity, enables LLMs to solve complex reasoning problems. Experiments show that our method achieves remarkable performance gains over single-stage fine-tuning, demonstrating the generalization on various tasks including logical, mathematical, and commensense reasoning tasks.

*Fenrong Liu and Robert van Rooij are corresponding authors.

2 Logical Question-Answering Tasks

Logical QA tasks are commonly used to assess the logical reasoning capabilities of LLMs, requiring LLMs to determine whether a statement can be derived from the given information. In logical QA tasks, there's a set of premises (denoted here as *facts*) and a statement (referred to as the *hypothesis*). The possible answers include: *Proved* (if it can be derived), *Disproved* (if it contradicts the facts), or *Unknown* (if the facts are insufficient to draw the hypothesis). An example for logical QA in dataset $\mathrm{FLD}_{\times 2}$ (Morishita et al. 2024) is as follows.

An example for Logical QA

Contexts (Facts):
- Someone that is dolomitic does prospect.
- If something does prospect it does benumb quartette.
- Something does not benumb quartette but it is a numeral if it does not prospect.
- .....

Query: Based on the provided facts, verify the hypothesis: That that opiate benumbs quartette hold.
Options: A) Proved B) Disproved C) Unknown
Answer: A) Proved

3 Proposed Method

Noting that LLMs' performance varies with the order of training samples during fine-tuning, we first quantify the logical reasoning complexity of training samples in symbolic formalization and then fine-tune the LLM on subsets (of the training data) in ascending order of the complexity.

Benchmark	Metric	LLaMA	LLaMA-Ours	Imp.	Ministral	Ministral-Ours	Imp.
LogiQA	acc	0.306 ± 0.018	$\mathbf{0.324 \pm 0.018}$	5.882% ↑	0.263 ± 0.017	$\mathbf{0.275 \pm 0.018}$	4.563% ↑
OpenBookQA	acc	0.313 ± 0.007	$\mathbf{0.315 \pm 0.007}$	0.639% ↑	0.336 ± 0.007	$\mathbf{0.349 \pm 0.007}$	3.869% ↑
HellaSwag	acc_norm	0.729 ± 0.004	$\mathbf{0.772 \pm 0.004}$	5.898% ↑	0.729 ± 0.004	$\mathbf{0.732 \pm 0.004}$	0.412% ↑
SciQ	acc_norm	0.949 ± 0.007	$\mathbf{0.954 \pm 0.007}$	0.527% ↑	0.919 ± 0.009	$\mathbf{0.921 \pm 0.009}$	0.218% ↑
GSM8k (CoT)	flexible-extract	0.616 ± 0.013	$\mathbf{0.626 \pm 0.013}$	1.623% ↑	0.074 ± 0.007	$\mathbf{0.144 \pm 0.010}$	94.595% ↑
GSM8k (CoT-Zero-Shot)	flexible-extract	0.454 ± 0.014	$\mathbf{0.466 \pm 0.014}$	2.643% ↑	0.055 ± 0.006	$\mathbf{0.063 \pm 0.007}$	14.545% ↑

Table 1: Performance evaluation comparing our reasoning complexity-based fine-tuning approach against single-stage fine-tuning across diverse reasoning benchmarks. Imp. indicates the relative improvement rate over the corresponding baseline.

Symbolic Formalization of Samples

To formalize the reasoning process, we first translate natural language inputs into symbolic language. We focus specifically on the propositional logic subset of the $\text{FLD}_{\times 2}$ dataset, denoted as $\mathcal{D}_{prop}$. Each sample in $\mathcal{D}_{prop}$ consists of a set of premises $\mathcal{P} = \{s_1, s_2, \ldots, s_n\}$, a hypothesis h and the label l. To execute translation, we employ a LLM, denoted as $\mathcal{M}_{trans}$. Specifically, $\mathcal{M}_{trans}$ first identifies atomic propositions and establishes a mapping dictionary (e.g., "Someone is dolomitic" assigns to p). Then the LLM reconstructs the logical structure in each premise with standard logical connectives ($\neg, \wedge, \vee, \rightarrow$). Consequently, the original input is converted into a symbolic set $\mathcal{P}_{sym} = \{\phi_1, \phi_2, \ldots, \phi_n\}$ and a target formula ψ, where ϕ_i and ψ are well-formed formulas (WFFs) in propositional logic. This formalization isolates the logical structure from vague natural language, facilitating precise logical reasoning complexity evaluation.

Reasoning Complexity Quantification of Samples

To quantify the logical complexity of each sample, we utilize DeepSeek as a reasoning complexity evaluator, denoted as $\mathcal{M}_{eval}$. For each sample $x_i \in \mathcal{D}_{prop}$, we prompt $\mathcal{M}_{eval}$ to generate a step-by-step deductive proof deriving the truth value of the hypothesis, given the label l_i.

Then we prompt $\mathcal{M}_{eval}$ to give a reasoning complexity score for each sample based on the proof path. Intuitively, problems requiring longer derivation chains or nested applications of inference rules (e.g., *Modus Tollens*) are assigned higher complexity scores. The complexity score function is denoted as $S(x_i)$. Based on scores, we sort the entire dataset $\mathcal{D}_{prop}$ in ascending order of reasoning complexity, obtaining a ranked sequence $\mathcal{D}_{sorted} = (x_{(1)}, x_{(2)}, \ldots, x_{(|\mathcal{D}_{prop}|)})$, where $S(x_{(i)}) \leq S(x_{(j)})$ for all $i < j$.

Complexity-Driven Dataset Stratification

Utilizing the ordered sequence $\mathcal{D}_{sorted}$, we stratify training dataset $\mathcal{D}_{prop}$ into several subsets. We divide $\mathcal{D}_{sorted}$ into three disjoint subsets: $\mathcal{D}_{easy}$, $\mathcal{D}_{medium}$, and $\mathcal{D}_{hard}$, based on the reasoning complexity scores. The partitions are defined by thresholds τ_1 and τ_2. This stratification ensures that $\mathcal{D}_{easy}$ contains samples with straightforward fewer deduction steps, while $\mathcal{D}_{hard}$ encompasses complex problems requiring longer reasoning steps and more logical rules.

Fine-tuning Based on Reasoning Complexity

Given $\mathcal{D}_{easy}$, $\mathcal{D}_{medium}$, and $\mathcal{D}_{hard}$, we further propose a phased fine-tuning framework that mimics the human learning process of progressing from simple to complex. We initialize our target LLM, $\mathcal{M}_\theta$, and update its parameters θ sequentially across three distinct training stages. In stage 1, the model is fine-tuned on $\mathcal{D}_{easy}$ to grasp fundamental logical rules, resulting in parameters θ_1. In stage 2, we fine-tune on $\mathcal{D}_{medium}$ to yield θ_2. Finally, in stage 3, the model with parameters θ_2 is trained on $\mathcal{D}_{hard}$ to obtain the final weights θ_{final}, which can tackle more complex reasoning patterns. This phased fine-tuning can help LLMs to tackle high-complexity propositional logical QA tasks.

4 Experiments

We fine-tune (Pang et al. 2024) two LLMs (LLaMA-3.1-8B (Dubey et al. 2024) and Ministral-8B (Team 2024)) on propositional logic subset of the $\text{FLD}_{\times 2}$, specifically utilizing samples partitioned by reasoning complexity. To assess reasoning performance, we employ five diverse reasoning benchmarks: LogiQA (Liu et al. 2021), OpenBookQA (Mihaylov et al. 2018), HellaSwag (Zellers et al. 2019), SciQ (Welbl, Liu, and Gardner 2017), and GSM8k (Cobbe et al. 2021). We adopt three metrics: *acc* denotes standard accuracy; *acc_norm* represents accuracy normalized by the length of the candidate answer; and *flexible-extract* validates whether the correct solution is present in the model's output.

Table 1 presents the comparative performance of our reasoning complexity-based fine-tuning approach against the single-stage baseline. Our method stably outperforms the baseline methods on all 5 benchmarks. Specifically, LLaMA-3.1-8B and Ministral-8B demonstrates remarkable gains with improvements of 5.882% on LogiQA and 94.595% on GSM8k (CoT), respectively. These results indicate that fine-tuning on samples ordered from lower to higher logical reasoning complexity effectively enhances the LLM's logical reasoning ability and shows robust generalization on mathematical and commonsense reasoning tasks.

5 Conclusion

In this work, we proposed a curriculum-based fine-tuning framework to enhance the logical reasoning capabilities of LLMs. By quantifying reasoning complexity of propositional logic samples and stratify the training data into three subsets, we fine-tune LLMs on these subsets ordered from easy to hard. Experimental results demonstrate that our method outperforms single-stage fine-tuning baselines across various reasoning benchmarks. For future work, we plan to extend this phased fine-tuning approach into broader logical domains, such as first-order logic.

Acknowledgments

FL was supported by the Beijing Natural Science Foundation (No. L257007) and Tsinghua University's Initiative for Advancing First-Class and World-Leading Disciplines in the Humanities and Social Sciences. RvR was supported by the EU-funded Marie Skłodowska-Curie Action (MSCA) PLEXUS project.

References

Cheng, F. 2025. Enhancing the logical reasoning abilities of large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 10969–10970.

Cheng, F. 2026. Empowering LLMs with Symbolic Representation and Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (Doctoral Consortuim)*.

Cheng, F.; Li, H.; Liu, F.; van Rooij, R.; Zhang, K.; and Lin, Z. 2025. Empowering LLMs with Logical Reasoning: A Comprehensive Survey. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, 10400–10408. Survey Track.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Feng, J.; Xu, R.; Hao, J.; Sharma, H.; Shen, Y.; Zhao, D.; and Chen, W. 2024. Language Models can be Deductive Solvers. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 4026–4042.

Fu, W.; Yang, H.; Cheng, F.; and Liu, F. 2025. Efficient First-Order Logic-Based Method for Enhancing Logical Reasoning Capabilities of LLMs. In *NeurIPS 2025 Workshop on Foundations of Reasoning in Language Models*.

Liu, J.; Cui, L.; Liu, H.; Huang, D.; Wang, Y.; and Zhang, Y. 2021. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 3622–3628.

Liu, T.; Xu, W.; Huang, W.; Zeng, Y.; Wang, J.; Wang, X.; Yang, H.; and Li, J. 2025. Logic-of-Thought: Injecting Logic into Contexts for Full Reasoning in Large Language Models. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 10168–10185.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2381–2391.

Morishita, T.; Morio, G.; Yamaguchi, A.; and Sogawa, Y. 2024. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. *Advances in Neural Information Processing Systems*, 37: 73572–73604.

Olausson, T.; Gu, A.; Lipkin, B.; Zhang, C.; Solar-Lezama, A.; Tenenbaum, J.; and Levy, R. 2023. LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5153–5176.

Pan, L.; Albalak, A.; Wang, X.; and Wang, W. 2023. Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3806–3824.

Pang, W.; Zhou, C.; Zhou, X.-H.; and Wang, X. 2024. Phased Instruction Fine-Tuning for Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 5735–5748.

Ryu, H.; Kim, G.; Lee, H. S.; and Yang, E. 2025. Divide and Translate: Compositional First-Order Logic Translation and Verification for Complex Logical Reasoning. In *The Thirteenth International Conference on Learning Representations*.

Team, M. A. 2024. Un Ministral, des Ministraux. https://mistral.ai/news/ministraux/.

Wan, Y.; Wang, W.; Yang, Y.; Yuan, Y.; Huang, J.-t.; He, P.; Jiao, W.; and Lyu, M. 2024. LogicAsker: Evaluating and Improving the Logical Reasoning Ability of Large Language Models. In *Proceed- ings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2124–2155.

Wang, X.; Yang, H.; Cheng, F.; and Liu, F. 2025. Adaptive Selection of Symbolic Languages for Improving LLM Logical Reasoning. In *AAAI 2026 Workshop on Post-AI Formal Methods*.

Welbl, J.; Liu, N. F.; and Gardner, M. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 94–106.

Xu, J.; Fei, H.; Luo, M.; Liu, Q.; Pan, L.; Wang, W. Y.; Nakov, P.; Lee, M.-L.; and Hsu, W. 2025. Aristotle: Mastering Logical Reasoning with A Logic-Complete Decompose-Search-Resolve Framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 3052–3075.

Xu, J.; Fei, H.; Pan, L.; Liu, Q.; Lee, M.-L.; and Hsu, W. 2024. Faithful Logical Reasoning via Symbolic Chain-of-Thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 10041–10058.

Ye, X.; Chen, Q.; Dillig, I.; and Durrett, G. 2023. Satlm: Satisfiability-aided language models using declarative prompting. *Advances in Neural Information Processing Systems*, 36: 45548–45580.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4791–4800.