Adaption-of-Thought: Learning Question Difficulty Improves Large Language Models for Reasoning

Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11), Tieyun Qian (/profile?id=~Tieyun_Qian1)

15 Jun 2024 (modified: 23 Aug 2024) ACL ARR 2024 June Submission June, Senior Area Chairs, Area Chairs, Reviewers, Authors, Ethics Reviewers, Ethics Chairs, Commitment Readers Revisions (/revisions?id=KVfYRQMgHI)

Abstract:
Large language models (LLMs) have shown excellent capability for solving reasoning problems. Existing approaches do not differentiate the question difficulty when designing prompting methods for them. Clearly, a simple method cannot elicit sufficient knowledge from LLMs to answer a hard question. Meanwhile, a sophisticated one will force the LLM to generate redundant or even inaccurate intermediate steps toward a simple question. Consequently, the performance of existing methods fluctuates among various questions.

In this work, we propose Adaption-of-Thought (\textsc{AdoT}), an adaptive method to improve LLMs for the reasoning problem, which first measures the question difficulty and then tailors demonstration set construction and difficulty-adapted retrieval strategies for the adaptive demonstration construction. Experimental results on three reasoning tasks prove the superiority of our proposed method, showing an absolute improvement of up to 5.5% on arithmetic reasoning, 7.4% on symbolic reasoning, and 2.3% on commonsense reasoning. Our codes and implementation details are available at: https://anonymous.4open.science/r/AdoT_anonymous (https://anonymous.4open.science/r/AdoT_anonymous)

Paper Type: Long
Research Area: Question Answering
Research Area Keywords: Reasoning, Large language models
Contribution Types: NLP engineering experiment, Approaches to low-resource settings
Languages Studied: English
Reassignment Request Action Editor: This is not a resubmission
Reassignment Request Reviewers: This is not a resubmission
A1 Limitations Section: This paper has a limitations section.
A2 Potential Risks: N/A
A3 Abstract And Introduction Summarize Claims: No
A3 Elaboration: Abstract; Section 1 Introduction
B Use Or Create Scientific Artifacts: No
B1 Cite Creators Of Artifacts: N/A
B2 Discuss The License For Artifacts: N/A
B3 Artifact Use Consistent With Intended Use: N/A
B4 Data Contains Personally Identifying Info Or Offensive Content: N/A
B5 Documentation Of Artifacts: N/A
B6 Statistics For Data: Yes
B6 Elaboration: Appendix C
C Computational Experiments: Yes
C1 Model Size And Budget: Yes
C1 Elaboration: Appendix G
C2 Experimental Setup And Hyperparameters: Yes
C2 Elaboration: Section 4.1; Appendix E;
C3 Descriptive Statistics: Yes
C3 Elaboration: Section 4.1
C4 Parameters For Packages: Yes
C4 Elaboration: Section 4.3; Appendix A
D Human Subjects Including Annotators: No

D1 Instructions Given To Participants: N/A
D2 Recruitment And Payment: N/A
D3 Data Consent: N/A
D4 Ethics Review Board Approval: N/A
D5 Characteristics Of Annotators: N/A
E Ai Assistants In Research Or Writing: Yes
E1 Information About Use Of Ai Assistants: Yes
E1 Elaboration: Section 4.1; Appendix H
Reviewing Volunteers: Tieyun Qian (/profile?id=~Tieyun_Qian1), Yongqi Li (/profile?id=~Yongqi_Li3)
Reviewing Volunteers For Emergency Reviewing: The volunteers listed above are only willing to serve as regular reviewers.
Reviewing No Volunteers Reason: N/A - An author was provided in the previous question.
Preprint: no
Preprint Status: There is no non-anonymous preprint and we do not intend to release one.
Consent To Share Data: yes
Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.
Author Submission Checklist: I confirm that the paper is anonymous and that all links to data/code repositories in the paper are anonymous., I confirm that the paper has proper length (Short papers: 4 content pages maximum, Long papers: 8 content pages maximum, Ethical considerations and Limitations do not count toward this limit), I confirm that the paper is properly formatted (Templates for *ACL conferences can be found here: https://github.com/acl-org/acl-style-files (https://github.com/acl-org/acl-style-files).)
Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I agree
Submission Number: 2136

Discussion (/forum?id=KVfYRQMgHI#discussion)

Filter by reply type... | Filter by author... | Search keywords... | Sort: Newest First

Everyone | Submission2136... | Submission2136 Area... | Submission2136 Authors | Submission2136... | Program Chairs | Submission2136... | Submission2136... | Ethics Chairs | Submission2136... | Submission2136... | Submission2136... | Submission2136...

25 / 25 replies shown

Add: **Author-Editor Confidential Comment** **Withdrawal**

Meta Review of Submission2136 by Area Chair hEpk

(https://openreview.net/forum?id=KVfYRQMgHI&noteId=OhczHWdBwi)

Meta Review by Area Chair hEpk 08 Aug 2024, 02:15 (modified: 23 Aug 2024, 07:02)
Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers
Revisions (/revisions?id=OhczHWdBwi)

Metareview:
The paper introduces a method called Adaption-of-Thought (ADOT) designed to enhance the reasoning capabilities of LLMs by considering the difficulty of questions. The method involves measuring question difficulty, tailoring demonstration sets, and applying difficulty-adapted retrieval strategies..

Summary Of Reasons To Publish:
ADOT is a novel solution that addresses the performance inconsistency of LLMs by aligning the complexity of prompts with the difficulty of questions. This innovative adaptation mechanism is both practical and effective.

Specifically, the enhancements include a 5.5% improvement in arithmetic reasoning, 7.4% in symbolic reasoning, and 2.3% in commonsense reasoning.

Summary Of Suggested Revisions:
Expand the evaluation to include more recent and powerful models (e.g., GPT-4).

Overall Assessment: 4 = There are minor points that may be revised

Suggested Venues: EMNLP Findings
Best Paper Ae: No
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Author Identity Guess: 1 = I do not have even an educated guess about author identity.

Add: **Author-Editor Confidential Comment**

Confidential Comment to Area Chair

Edit

Author-Editor Confidential Comment
by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))
03 Aug 2024, 16:46 Program Chairs, Senior Area Chairs, Area Chairs, Authors

Comment:
Dear Area Chair,

First, we genuinely appreciate your dedicated efforts in organizing the conference! We also appreciate all reviewers' comments and suggestions. We have presented detailed responses to address the reviewers' concerns. However, it has been nearly 5 days without further discussion or queries from the reviewer abst and 6jWX.

(1) The reviewer 6jWX comments that "This is, in general, a good work, although there are some missing details and concerns over the generalization capability of the approach. I am giving a conservative score due to missing details but will raise the score if they are clarified." and "How is step decomposition implemented? This part is important but seems to be missing from the main script. I will raise the score if this part is resolved".

Regarding the detailed implementation of step decomposition, we have meticulously introduced it in our response. Furthermore, all implementations of our method (including data, data preprocessing programs, and code implementations for every part of our method) had been public in the anonymous GitHub repository linked on the first page of our original paper before we submitted it.

As for the generalization capability of the approach, some supplement experimental results show that our method is competitive compared to the sota baseline.

(2) The reviewer abst mainly concerned with the difficulty measurement and distribution, we have posted detailed responses to alleviate the concerns.

We will sincerely appreciate your kind justification of our paper, considering the paper content and our detailed responses. In particular, reviewer 6jWX kindly mentioned twice that he/she will raise the score if the implementation of step decomposition is clarified, but after providing the details in responses, we cannot get any feedback even sending three reminders.

We greatly appreciate your time and effort in our work!

Best regards!

Paper2136 authors.

Add: **Author-Editor Confidential Comment**

Official Review of Submission2136 by Reviewer abst

Official Review by Reviewer abst 21 Jul 2024, 11:07 (modified: 23 Aug 2024, 07:02)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer abst, Commitment Readers
Revisions (/revisions?id=Ii8fuLOrQc)

Paper Summary:
The paper introduces a method called Adaption-of-Thought (ADOT) that aims to improve large language models (LLMs) in solving reasoning problems by considering the difficulty of questions. Existing methods do not account for question difficulty, leading to performance inconsistencies. ADOT measures question difficulty and adapts demonstration sets and

retrieval strategies accordingly, showing significant performance improvements across arithmetic, symbolic, and commonsense reasoning tasks.

Summary Of Strengths:

The paper introduces the Adaption-of-Thought (ADOT) method, which tailors the complexity of prompts to match the difficulty of questions. Key features include:

Measuring Question Difficulty: Evaluates syntactic and semantic complexities.

Adaptive Demonstration Set Construction: Creates demonstration sets that align with question difficulty.

Difficulty-Adapted Retrieval: Selects demonstrations that closely match the target question's difficulty.

2. Significant Improvement

ADOT demonstrates substantial performance enhancements across multiple reasoning tasks These improvements underscore the effectiveness of adjusting prompts based on question difficulty.

3. Comprehensive Analysis The paper offers a detailed evaluation of ADOT:

Ablation Studies: Validates the necessity of each component.

Task-Specific Evaluations: Tests across multiple datasets and tasks.' Comparison with State-of-the-Art Methods: Shows consistent superiority over thirteen existing methods.

Supplemental Experiments: Includes tests on computational efficiency, generalization across LLM sizes, and sensitivity to different demonstrations.

Summary Of Weaknesses:

Definition of difficulty in the paper: the paper uses two methods to calculate the difficulty -- syntactic complexity and semantic complexity. For syntactic complexity, the authors use the length of the rationale as syntactic complexity. For semantics complexity, the authors calculate how many different tokens are between the question and rationale. Firstly, I think there could be more methods to calculate those two complexities. Second, whether those two criteria can reflect question difficulty is uncertain to me. Can authors provide some statistics to show that your criteria are actually aligned with question difficulty distribution?

Comments Suggestions And Typos:

1. line 255, wrong use of citation
2. In the first part: measuring the difficulty, what if the rationale itself is wrong? How can you make sure the rationale reflects the difficulty? did you do any manual checks of rationale quality?

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.
Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.
Best Paper: No
Needs Ethics Review: No
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 1 = No usable datasets submitted.
Software: 1 = No usable software released.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Add: **Author-Editor Confidential Comment**

Kindly Reminder to Reviewer abst

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

02 Aug 2024, 19:37 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Authors, Ethics Reviewers, Ethics Chairs, Reviewer abst, Commitment Readers

Revisions (/revisions?id=P3SKAE3bsT)

Comment:
Dear Reviewer,

Due to your busy schedule, we know that you may not have sufficient time to see our responses. However, as the current rebuttal phase is coming to an end, we have to send this reminder. Since we posted detailed responses to alleviate your concerns four days ago, we have not received your feedback yet. Could you kindly check our rebuttal and let us know if our responses have addressed your concerns? This would be very important for improving our work. Thank you once again.

Best wishes!

Paper2136 authors.

Add: **Author-Editor Confidential Comment**

Kindly Reminder to Reviewer abst

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

01 Aug 2024, 16:38 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer abst, Commitment Readers

Revisions (/revisions?id=YJocPD4n8Y)

Comment:
Dear reviewer,

We are sorry to bother you again since the current rebuttal phase is coming to an end. Two reviewers who rate our work as solid work and give an overall 4 score have responded to us. After discussion, they believe the 4 score is appropriate. We really appreciate their valuable feedback.

We have posted detailed responses to resolve your concerns about difficulty measurement and statistics of difficulty distribution. We wonder what is your opinion about our response. We are looking forward to receiving feedback from you and sincerely appreciate your time and effort.

Best wishes!

Paper2136 authors.

Add: **Author-Editor Confidential Comment**

Kindly Requesting Your Valuable Feedback.

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

31 Jul 2024, 17:05 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer abst, Commitment Readers

Revisions (/revisions?id=pkKeVAxvO4)

Comment:
Dear reviewer,

We are sorry to bother you. Could you kindly please take a look at our responses? We wonder if we have addressed your concerns. If you have any further questions, please don't hesitate to tell us, so we can make a response again before the rebuttal deadline.

We greatly appreciate your dedicated time and effort in our work. Your feedback is very important to us. We are looking forward to communicating with you further!

Best regards!

Paper2136 authors.

Add: **Author-Editor Confidential Comment**

Responses to Reviewer abst (part 1)

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

29 Jul 2024, 21:44 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer abst, Commitment Readers

Revisions (/revisions?id=yGhZ3vnp9i)

Comment:

Dear reviewer,

Thank you for your insightful suggestions. Here are the responses to address your concerns.

W1: Definition of difficulty in the paper: the paper uses two methods to calculate the difficulty -- syntactic complexity and semantic complexity. For syntactic complexity, the authors use the length of the rationale as syntactic complexity. For semantics complexity, the authors calculate how many different tokens are between the question and rationale. Firstly, I think there could be more methods to calculate those two complexities.

R1: As you have pointed out, there exist more methods to calculate those two complexities. In our preliminary research, we designed multiple methods for calculating these two types of complexity. For example, adopting some features of the Syntactic Dependency Tree and Semantic Dependency Tree of the rationale as measures for two complexities, such as the tree's diameter and depth (positively related to the syntactic and semantic structure complexity). However, compared to directly using the length and additional semantic words, these methods do not yield significant improvements. Meanwhile, these methods reduce computational efficiency because they require additional tools (such as Stanford CoreNLP, Spacy, and Networkx) to perform complex processing on the rationales. Therefore, to make our approach more practical, we ultimately adopt the current method. As shown in Appendix G Computational Efficiency (pages 14-15, lines 1079-1127), our model can maintain high computational efficiency and improve reasoning performance simultaneously when using the current complexity measurement method.

W2: Second, whether those two criteria can reflect question difficulty is uncertain to me. Can authors provide some statistics to show that your criteria are actually aligned with question difficulty distribution?

R2: Yes, below are some statistics and explanations. To prove that our criteria are actually aligned with question difficulty distribution, we designed the following experiment. First, we sort all the questions in a descending order of the calculated difficulty, and then divide them into 10 sections with equal question numbers. The accuracy of each section can reflect the actual difficulty. Specifically, the lower the accuracy, the higher the actual difficulty. As shown in Supplement Table 1 and 2, when the calculated difficulty increases, the accuracy generally decreases, denoting that the actual difficulty increases. The statistics result proves that our criteria are actually aligned with question difficulty distribution.

Difficulty section	**15-42**	**43-52**	**53-61**	**62-71**	**71-88**	**88-102**	**103-126**	**126-142**	**143-176**	**179-283**
Accuracy (%)	80.0	80.0	72.0	68.0	56.0	52.0	56.0	48.0	32.0	28.0

Supplement Table 1: Accuracy (%) on 10 difficulty sections on the AQuA dataset.

Difficulty section	**22-46**	**46-54**	**54-61**	**61-67**	**67-73**	**73-81**	**81-88**	**88-101**	**101-119**	**119-200**
Accuracy (%)	90.1	86.3	84.7	79.4	80.9	80.2	74.8	75.6	67.9	61.8

Supplement Table 2: Accuracy (%) on 10 difficulty sections on the GSM8K dataset.

Add: **Author-Editor Confidential Comment**

Responses to Reviewer abst (part 2)

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

29 Jul 2024, 21:46 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer abst, Commitment Readers

Revisions (/revisions?id=GXBtTp7nZP)

Comment:

S1: line 255, wrong use of citation

R1: We sincerely apologize for the incorrect use of citation. We have identified 3 potential issues with this citation: (1) (Wei et al., 2022) → Wei et al. (2022)

(2) We originally added the citation information based on Google Scholar, but upon comparison with the ACM Digital Library, we discovered that one author's information was missing. This author was added when the paper was officially published on April 3, 2024.

Google Scholar citation： Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837.

ACM Digital Library citation： Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, **Brian Ichter**, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 1800, 24824–24837.

(3) We clarify the specific location of the original demonstrations in the original paper: we propose to not only improve the quality of the original demonstrations in **Appendix G: Full Prompts** of Wei et al. (2022)...

We are very grateful for your reminder. Additionally, we have conducted a thorough review of all citations throughout the entire paper to ensure that no wrong citations will appear in the next version.

S2: In the first part: measuring the difficulty, (1) what if the rationale itself is wrong? (2) How can you make sure the rationale reflects the difficulty? (3) did you do any manual checks of rationale quality?

R2: (1) To the first question: We believe that even if this rationale contains some errors, it will not significantly impact the measurement of difficulty. Our analysis is shown as follows:

In practice, difficulty is a relative concept. Whether a question is easy or hard is referenced against other questions. Note that we aim to measure the relative difficulty between different questions, not the absolute difficulty. To this end, we use the same LLMs and demonstrations, under completely consistent environments, to generate rationales for each question. We then measure difficulty based on these rationales. Even if there are errors in the rationales, the only variable in the "(LLMs + demonstrations + question + environments)→rationale→difficulty" pipeline is the question itself. Therefore, using this method to measure the relative difficulty between different questions is fair and reliable.

(2) To the second question: Rationale is the reasoning process of a question. Generally, the more complex the reasoning process, the more hard the question usually is. Hence, the rationale can reflect the difficulty of question.

(3) To the third question: We randomly select 50 rationales from the AQuA dataset. Then, one PhD student and two Master student who are familiar with this task manually check their quality. After discussion, we reached the following consensus on the quality of rationales: First, all rationales are reasonable in linguistic expression, without meaningless and off-topic text. Second, all rationales are highly consistent in format because they are generated by the same LLM, demonstrations, and environment. Third, compared with correct rationales, incorrect rationales are usually longer and more difficult to understand.

Thank you very much for your valuable suggestions again. We hope our responses can address your concerns. We also sincerely hope these efforts can improve the soundness and overall assessment of our work. If you have any further comments or suggestions, please do not hesitate to let us know.

Best regards!

Paper 2136 authors.

Add: **Author-Editor Confidential Comment**

Official Review of Submission2136 by Reviewer pde1

Official Review by Reviewer pde1 21 Jul 2024, 06:05 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer pde1, Commitment Readers

Revisions (/revisions?id=IbJJWLL2gj)

Paper Summary:
The paper proposes a framework for adaptively selecting demonstrations based on the difficulty of the questions. The authors first point out that none of the state-of-the-art methods performs consistently best for different types of QA benchmark datasets. They then propose a new framework, called Adaption-of-Thought, which consists of the following three steps: (1) measure question difficulty, (2) adaptively build a demonstration set, (3) adaptively retrieve the demonstration. The effectiveness of the proposed method is demonstrated by experimenting with 10 QA benchmark datasets.

Summary Of Strengths:
- The paper is well written with clear motivation and scenario which could be a good reference for related researchers.
- The proposed adaptive prompting method showed significant performance improvement over existing prompting strategies.

Summary Of Weaknesses:
- The evaluation is done only for a single LLM (gpt-3.5-turbo-0613). A small investigation to see how it works with recent models would be helpful.

Comments Suggestions And Typos:
- Investigating transferability (generalizability) of the difficlty measure would be interesting. Similarity-based might be costless and robust baseline.
- It would be interesting to further elaborate on the following issues. Note that these are the strengths of this paper in that they provide a starting point for such a discussion.
  - The definition of question difficulty is important in this framework. At the moment, the formulas (2)(3)(4) are rather naive and could be deepened in the discussion.
  - The easy, normal, and hard levels of difficulty are discussed in terms of coreference and complicated reasoning steps, but whether this is sufficient is also an important research issue. In addition, the definition of problem difficulty includes many research questions, such as coreference is not only a pronoun problem, and decomposition is also a challenging problem for LLMs.

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.
Soundness: 3.5
Overall Assessment: 4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it.
Best Paper: No
Needs Ethics Review: No
Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
Datasets: 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.
Software: 3 = Potentially useful: Someone might find the new software useful for their work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Add: **Author-Editor Confidential Comment**

Responses to Reviewer pde1 (part 1)

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

29 Jul 2024, 21:47 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer pde1, Commitment Readers

Revisions (/revisions?id=7NsgUTDOhg)

Comment:

Dear reviewer,

Thank you for your insightful suggestions. Here are the responses to address your concerns.

W1: The evaluation is done only for a single LLM (gpt-3.5-turbo-0613). A small investigation to see how it works with recent models would be helpful.

R1: Besides gpt-3.5-turbo-0613, we have conducted experiments on llama2-7b-chat, llama2-13b-chat, llama2-70b-chat (Appendix H: Method Generalization on LLMs with Different Sizes, Pages 15-16), which vary from different sizes. As these experimental results show, our method can achieve the best results on a wide range of LLMs in most cases.

To further address your concerns, we will do our best to complete the experiments on more recent gpt-4. Due to the high price of gpt-4 (60 times gpt-3.5-turbo-0613/per token) and our limited budget, we choose the Top-5 best methods and the most commonly-used datasets in each task. The experimental results are shown in Supplement Table 3, which proves that our method still can achieve the sota performance with an average absolute improvement of 2.9%.

Task	***Arithmetic***	***Symbolic***	***Commonsense***	
Dataset	**AQuA**	**Last Letters**	**CSQA**	**Average**
Few-shot-cot	70.1	95.4	84.3	83.3
Complex-cot	69.3	<u>96.2</u>	84.6	83.4
Self-ask	70.5	94.6	83.8	83.0
Least-to-most	68.5	95.8	84.4	82.9
PHP	<u>72.8</u>	94.8	<u>84.8</u>	<u>84.1</u>
AdoT(ours)	**77.2**	**97.4**	**86.4**	**87.0**

Supplement Table 3: performance comparison on gpt-4 (%) (**bold:** the best score, <u>underline</u>: the second best score).

S1: Investigating transferability (generalizability) of the difficulty measure would be interesting. Similarity-based might be costless and robust baseline.

R1: In section 4.5: Effectiveness of Retrieval Method (pages 7-8, lines 521-545), we have conducted similar experiments. Specifically, we design a similarity-based method, which leverages text similarity as another type of difficulty measure. This method retrieves some demonstrations where the sample questions have the closest text similarity to the target questions. As you point out, the similarity-based method is robust with high performance. Furthermore, its generalizability in three tasks is also excellent. However, compared to the similarity-based method, our method performs better.

Add: **Author-Editor Confidential Comment**

Responses to Reviewer pde1 (part 2)

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

29 Jul 2024, 21:47 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer pde1, Commitment Readers

Revisions (/revisions?id=j9hpiFeQom)

Comment:

S2: It would be interesting to further elaborate on the following issues. Note that these are the strengths of this paper in that they provide a starting point for such a discussion.

- (1) The definition of question difficulty is important in this framework. At the moment, the formulas (2)(3)(4) are rather naive and could be deepened in the discussion.

- (2) The easy, normal, and hard levels of difficulty are discussed in terms of coreference and complicated reasoning steps, but whether this is sufficient is also an important research issue. In addition, the definition of problem difficulty includes many research questions, such as coreference is not only a pronoun problem, and decomposition is also a challenging problem for LLMs.

R2: Many thanks you for your positive feedback on our paper!

(1) For the first point, in our preliminary research, we designed multiple methods for measuring difficulty. For example, adopting some features of the Syntactic Dependency Tree and Semantic Dependency Tree of the rationale as measures for difficulty, such as the tree's diameter and depth (positively related to the syntactic and semantic structure complexity). However, compared to directly using the length and additional semantic words, these methods do not yield significant improvements. Meanwhile, these methods reduce computational efficiency because they require additional tools (such as Stanford CoreNLP, Spacy, and Networkx) to perform complex processing on the rationales. Therefore, to make our approach more practical, we ultimately adopt the current method. As shown in Appendix G Computational Efficiency (pages 14-15, lines 1079-1127), our model can maintain high computational efficiency and improve reasoning performance simultaneously when using the current complexity measurement method.

(2) For the second point, first of all, we must thank you for your valuable questions, which greatly inspires our future work. We would like to discuss it with you from the following points of view.

(a) First, as you pointed out, other problems may exist in different difficulty sections besides coreference and complicated reasoning steps. For instance, some unfamiliar words may exist in the rationales of hard questions, which also prevents the LLMs from understanding them.

(b) Second, as you pointed out, coreference includes not only pronoun coreference, but also noun phrase coreference, event coreference, etc. We use a tool [1] to complete this process, which is mainly good at resolving pronoun and noun references. The goal of the step decomposition is to let the LLM know how to decompose a complex rationale into an easy-to-understand rationale by adding more steps. To this end, we show LLM a pair of (a compact rationale of the original rationale, the original rationale). The compact rationale is formed by the step deletion, which is the reverse process of step decomposition. Specifically, given a rationale with N steps, we randomly delete $\frac{N-1}{2}$ steps chosen from the first step to the $(N-1)$-th step, where the last step should be kept because it is often the final answer. By imitating the process in $\left[\text{a compact rationale of the original rationale} \overset{step\ decomposition}{\longrightarrow} \text{the original rationale}\right]$, LLMs will conduct $\left[\text{the original rationale} \overset{step\ decomposition}{\longrightarrow} \text{a decomposed rationale of the original rationale}\right]$.

In future works, we will continue to explore these question based on your insightful suggestion.

Reference: [1] https://github.com/huggingface/neuralcoref (https://github.com/huggingface/neuralcoref)

Thank you very much for your valuable suggestions again. We hope our responses address your concerns.

Best regards!

Paper 2136 authors.

Add: **Author-Editor Confidential Comment**

Replying to Responses to Reviewer pde1 (part 2)

Thank you for your response

Official Comment by Reviewer pde1 01 Aug 2024, 07:28 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer pde1, Commitment Readers

Revisions (/revisions?id=MWIjcfqluJ)

Comment:

Thank you for your response. These discussions are helpful for other researchers to further investigate the topic. (I have already taken this into account in my rating, so the score remains the same).

Checking the effectiveness with more powerful LLMs might clarify the target LLMs of the method (if it is mainly for Llama 2 class or smaller LLMs, it is still ok, but such information itself is valuable).

Add: **Author-Editor Confidential Comment**

Replying to Thank you for your response

Official Comment by Authors

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

01 Aug 2024, 09:13 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer pde1, Commitment Readers

Revisions (/revisions?id=xeiaQcbRmC)

Comment:

Dear reviewer,

Thank you so much for taking the time to read our responses. We will add these discussions to the paper based on your valuable suggestions. Thanks again.

Best wishes!

Paper2136 authors.

Add: **Author-Editor Confidential Comment**

Official Review of Submission2136 by Reviewer xZdU

Official Review by Reviewer xZdU 20 Jul 2024, 11:14 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer xZdU, Commitment Readers

Revisions (/revisions?id=LmQBnU0Wo6)

Paper Summary:

Large language models (LLMs) excel at reasoning but struggle with question difficulty. Existing methods are either too simple for hard questions or too complex for easy ones, leading to inconsistent performance. The Adaption-of-Thought (ADOT) method addresses this by assessing question difficulty and adapting prompts accordingly. ADOT improves performance, with gains of up to 5.5% in arithmetic reasoning, 7.4% in symbolic reasoning, and 2.3% in commonsense reasoning.

Summary Of Strengths:

1. Examples using the AQUA dataset and the overall framework are well expressed in an easy-to-understand manner.
2. Through a case study, the results for the sample question were shown.
3. Through comparison with various methods based on various benchmark datasets, performance improvement of the proposed method was shown.

Summary Of Weaknesses:

1. Did the method not consider the difficulty of the word itself (advanced vocabulary)?
2. Does difficulty correspond to the difficulty that most people actually feel? If not, is the scale for calculating difficulty proposed in the paper more accurate or reliable than one done by humans?
3. In step 2, section partitioning is performed based on the difficulty of the sampled questions. If the randomly sampled questions are felt by people to be very difficult overall compared to other questions, they are categorized as 'easy', 'normal', How is it divided into 'hard'?
4. For example, when the 'easy' part uses the same process as the 'normal' part or the 'normal' part is set to 'hard' ('hard', 'normal' OR 'hard', 'easy') in the paper. Is the performance improvement greater than the proposed method ('easy', 'normal', 'hard')?

Comments Suggestions And Typos:

See the weaknesses section above.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.
Overall Assessment: 4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it.
Best Paper: No
Limitations And Societal Impact:
Yes, the author has appropriately discussed the limitations and potential positive and negative social impacts of his work.

Ethical Concerns:
None

Needs Ethics Review: No
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 3 = Potentially useful: Someone might find the new datasets useful for their work.
Software: 3 = Potentially useful: Someone might find the new software useful for their work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Additional: No

Add: **Author-Editor Confidential Comment**

Responses to Reviewer xZdU (part 1)

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

29 Jul 2024, 21:49 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer xZdU, Commitment Readers

Revisions (/revisions?id=EWfacimKZk)

Comment:
Dear reviewer,

Thank you for your valuable suggestions. Here are the responses to address your concerns.

W1: Did the method not consider the difficulty of the word itself (advanced vocabulary)?

R1: Thank you very much for the constructive suggestions, which inspired us to conduct a more in-depth study around the word itself as follows:

First, we conducted a statistical analysis of the unfamiliar vocabulary and advanced vocabulary within the rationale.

Regarding unfamiliar vocabulary, a rationale containing more unfamiliar vocabulary indicates that it introduces more rare concepts in the reasoning process, making the problem more difficult to solve. Specifically, we base our analysis on word frequency in Wikipedia[1], categorizing the top 10% of words as common words and the remaining 90% as unfamiliar vocabulary. The results of the unfamiliar vocabulary statistics in the rationale are shown in the second row of Supplement Table 4. Clearly, the unfamiliar vocabulary ratio is extremely low. For example, in the AQuA dataset, there are only 16 instances of unfamiliar vocabulary across all rationales (including many repeated occurrences). We find that the characteristic of unfamiliar vocabulary in reasoning tasks is not significant, making it challenging to apply it to distinguish the difficulty of reasoning problems.

Regarding advanced vocabulary, a rationale containing more advanced vocabulary indicates that more advanced concepts are introduced in the reasoning process, requiring deeper knowledge to solve the problem. Specifically, we base our analysis on a previously constructed advanced vocabulary list[2], built from 15 different sources of advanced vocabulary collections. The results of advanced vocabulary ratio statistics in the rationale are shown in the third row of Supplement Table 4. Similar to the unfamiliar vocabulary ratio, the advanced vocabulary ratio is also extremely low, with no more than 10 occurrences in some datasets (such as Last Letters and Coin Flip). Therefore, similar to unfamiliar vocabulary, we find that it is challenging to extract effective features of advanced vocabulary in reasoning problems.

Although it is currently difficult to extract word-level difficulty, as you insightfully pointed out, this remains a very worthy area of exploration. We will supplement the discussion on this part in future versions to provide insights for related work in the NLP community.

Dataset	AQuA	GSM8K	SVAMP	AddSub	MultiArith	SingleEQ	Last Letters	Coin Flip	CSQA	StrategyQA
unfamiliar vocabulary ratio (%)	0.068	0.422	0.870	0.202	0.287	0.294	1.850	0.100	0.071	0.429
advanced vocabulary ratio (%)	0.195	0.135	0.157	0.290	0.132	0.138	0.033	0.005	0.300	0.646

Supplement Table 4: unfamiliar vocabulary ratio and advanced vocabulary ratio

Reference:

[1] https://github.com/IlyaSemenov/wikipedia-word-frequency (https://github.com/IlyaSemenov/wikipedia-word-frequency)

[2] https://github.com/Isomorpheuss/advanced-english-vocabulary (https://github.com/Isomorpheuss/advanced-english-vocabulary)

Add: **Author-Editor Confidential Comment**

Responses to Reviewer xZdU (part 2)

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

29 Jul 2024, 21:49 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer xZdU, Commitment Readers

Revisions (/revisions?id=LKLZcSbsKl)

Comment:

W2: (1) Does difficulty correspond to the difficulty that most people actually feel? (2) If not, is the scale for calculating difficulty proposed in the paper more accurate or reliable than one done by humans?

R2: (1) For the first question: To explore this issue, we first randomly select 50 questions from AQuA and Commonsenseqa datasets, respectively. Then, a PhD student and two Master students who are familiar with the task manually judge the question difficulty following these principles：

Easy question: the person can answer it without hesitation in 30 seconds.

Normal question: the person can answer it without hesitation in 120 seconds.

Hard question: the person cannot answer it in 120 seconds.

The final difficulty is obtained by voting on the manual judgment results of three persons. Meanwhile, when three persons' judgments were inconsistent with each other, we asked another PhD student to judge the final result. We use a confusion matrix to present the difference between manual judgment difficulty and automatically calculated difficulty. For instance, in Supplement Table 5, there exist 2 questions that most people feel are easy but are calculated as hard.

	easy (calculated)	**normal **(calculated)	hard (calculated)
easy (manual)	10	5	2
normal (manual)	3	14	8
hard (manual)	0	3	5

Supplement Table 5: the confusion matrix between manual judgment difficulty and automatically calculated difficulty in the AQuA dataset.

	easy (calculated)	**normal **(calculated)	hard (calculated)
easy (manual)	2	40	2
normal (manual)	0	4	0
hard (manual)	0	2	0

Supplement Table 6: the confusion matrix between manual judgment difficulty and automatically calculated difficulty in the CommonsenseQA dataset.

From the results in Tables 5 and 6, the manual judgment difficulty and automatically calculated difficulty are inconsistent in many cases.

(2) For the second question: In the AQuA dataset, 20% of the questions have completely inconsistent judgments from three persons, 86% have at least two judgments, and only 14% have three same judgments. In the CommonsenseQA dataset, 10% of the questions have completely inconsistent judgments from three persons, 44% have at least two judgments, and 56% have three same judgments.

This shows that relying on manual judgment is extremely unstable and unreliable. In addition, relying on manual judgment is not practical enough and is difficult to deploy in reality.

Other responses: This is a very valuable question and worth to explore. The authors have conducted an in-depth discussion on this issue. Here is the consensus we have reached:

When measuring difficulty, it is essential to first identify the subject. Specifically, it is important to clarify whether the difficulty is for people or for LLMs, which makes a difference. For example, previous research [1] has shown that some specific questions are very simple for humans but cannot be answered correctly by LLMs. At the same time, LLMs perform better than humans in solving certain specific tasks [2]. Therefore, it is challenging to define the absolute difficulty of a question. A question may be easy for humans but hard for LLMs. Conversely, the opposite situation can also occur. Since our goal is to explore how to utilize LLMs to answer questions, we need to measure difficulty from the perspective of LLMs. Therefore, whether the difficulty aligns with the difficulty of the human perspective is not the focus of our study on LLMs reasoning.

Reference:

[1] Berglund L, Tong M, Kaufmann M, et al. The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”[C]//The Twelfth International Conference on Learning Representations (2024).

[2] Pu X, Gao M, Wan X. Summarization is (almost) dead[J]. arXiv preprint arXiv:2309.09558, 2023.

Add: **Author-Editor Confidential Comment**

Replying to Responses to Reviewer xZdU (part 2)

Responses to Reviewer xZdU (part 3)

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

29 Jul 2024, 21:50 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer xZdU, Commitment Readers

Revisions (/revisions?id=4grDcq62br)

Comment:

W3: In step 2, section partitioning is performed based on the difficulty of the sampled questions. If the randomly sampled questions are felt by people to be very difficult overall compared to other questions, they are categorized as 'easy', 'normal', How is it divided into 'hard'?

R3: This is indeed a good question. In fact, we have the same question when we design the method. After in-depth analysis and verification, we found that when the number of randomly sampled questions is not extremely small (such as no more than 30), the distribution of the randomly sampled questions is close to the distribution of the sampling source. Therefore, there will not be a situation where only hard questions are sampled.

We can use a set of data to prove this. We have 1000 easy, normal, and hard questions, respectively, and we randomly sample 50 questions from these 3000 questions and repeat this process 5 times. The sampling results of [easy: normal: hard] are shown as follows:

[18,17,15]

[16,19,15]

[13,19,17]

[18,18,14]

[19,15,16]

From the above results, we can observe that these distributions are close to the original distributions, and there will be no situation where only a certain type of question is sampled.

W4: For example, when the 'easy' part uses the same process as the 'normal' part or the 'normal' part is set to 'hard' ('hard', 'normal' OR 'hard', 'easy') in the paper. Is the performance improvement greater than the proposed method ('easy', 'normal', 'hard')?

R4: We have conducted a similar experiment in the Appendix F: The Effectiveness of Coreference Resolution and Step Decomposition (pages 13-14, lines 1013-1078). Specifically, each part generally achieves the best results when its corresponding process is employed. It is important to note that using the hard process for easy part and the easy process for hard part will lead to the poorest performance.

Thank you very much for your valuable suggestions again. We hope our responses can address your concerns.

Best regards!

Paper 2136 authors.

Add: **Author-Editor Confidential Comment**

Replying to Responses to Reviewer xZdU (part 2)

Thank you for your response

Official Comment by Reviewer xZdU 01 Aug 2024, 09:39 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer xZdU, Commitment Readers

Revisions (/revisions?id=5isTr7jlj8)

Comment:

Thank you for your response. I have carefully reviewed the sincere responses to my comments. Even though the score mentioned have been reflected, I still believe the given score is appropriate.

Add: **Author-Editor Confidential Comment**

Replying to Thank you for your response

Official Comment by Authors

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

01 Aug 2024, 10:05 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer xZdU, Commitment Readers

Revisions (/revisions?id=m26QqoHduo)

Comment:

Dear reviewer,

Thank you so much for taking the time to read our responses. We will add these discussions to the paper based on your insightful suggestions. Thanks again.

Best wishes!

Paper2136 authors.

Add: **Author-Editor Confidential Comment**

Official Review of Submission2136 by Reviewer 6jWX

Official Review by Reviewer 6jWX 19 Jul 2024, 08:54 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer 6jWX, Commitment Readers

Revisions (/revisions?id=r9xL0aqerj)

Paper Summary:
This work proposes to retrieve demonstration examples of different difficulties and answer complexities to adaptively adjust the reasoning behavior of language models. In principle, the approach enables shorter answers for easy questions to avoid inaccuracies in lengthy answers, while ensure sufficiently long answers for more complex questions. The highlight is that this work uses demonstration examples to modulate the model behavior, instead of using natural language prompts. The proposed pipeline demonstrates consistent improvements over baselines. This is in general a good work, although there are some missing details and concerns over the generalization capability of the approach. I am giving a conservative score due to missing details but will raise the score if they are clarified.

Summary Of Strengths:
- The proposed approach is clear in a higher level
- The performance improvement is consistent although small
- The choice of using demonstration samples rather than natural language prompts to control the reasoning length is very reasonable

Summary Of Weaknesses:
- A demonstration pool needs to be established in advance. This will not only increase the cost of the method, but also limit the generalization capability. It is not clear about the model performance on out-of-distribution questions.
- How is step decomposition implemented? This part is important but seems to be missing from the main script. I will raise the score if this part is resolved.

Comments Suggestions And Typos:
- line 169, use $i - 1$ instead of i - 1 (include -1 inside the inline equation)

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.
Overall Assessment: 2.5
Best Paper: No
Needs Ethics Review: No
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 1 = No usable datasets submitted.
Software: 3 = Potentially useful: Someone might find the new software useful for their work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Add: **Author-Editor Confidential Comment**

Kindly Reminder to Reviewer 6jWX

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

02 Aug 2024, 19:38 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Authors, Ethics Reviewers, Ethics Chairs, Reviewer 6jWX, Commitment Readers

Revisions (/revisions?id=mTxi2s5fAL)

Comment:

Dear Reviewer,

Due to your busy schedule, we know that you may not have sufficient time to see our responses. However, as the current rebuttal phase is coming to an end, we have to send this reminder. Since we posted detailed responses to alleviate your concerns four days ago, we have not received your feedback yet. Could you kindly check our rebuttal and let us know if our responses have addressed your concerns? You kindly mentioned twice that you will raise the score if the implementation of step decomposition is clarified. This would be very important for improving our work. Thank you once again.

Best wishes!

Paper2136 authors.

Add: **Author-Editor Confidential Comment**

Kindly Reminder to Reviewer 6jWX

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

01 Aug 2024, 16:34 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer 6jWX, Commitment Readers

Revisions (/revisions?id=cERniPUvEO)

Comment:

Dear reviewer,

We are sorry to bother you again since the current rebuttal phase is coming to an end. Two reviewers who rate our work as solid work and give an overall 4 score have responded to us. After discussion, they believe the 4 score is appropriate. We really appreciate their valuable feedback. We are also encouraged that you evaluate our work as good work!

We have posted detailed responses to resolve your concerns about computational efficiency, generalization capability, and the implementation of step decomposition. We genuinely hope that you can reconsider the overall rating since you kindly mentioned that you will raise the score if the implementation of step decomposition is clarified. We are looking forward to receiving feedback from you and sincerely appreciate your time and effort.

Best wishes!

Paper2136 authors.

Add: **Author-Editor Confidential Comment**

Kindly Requesting Your Valuable Feedback.

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

31 Jul 2024, 17:06 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer 6jWX, Commitment Readers

Revisions (/revisions?id=p17YhGL7By)

Comment:

Dear reviewer,

We are sorry to bother you. Could you kindly please take a look at our responses? We wonder if we have addressed your concerns. If you have any further questions, please don't hesitate to tell us, so we can make a response again before the rebuttal deadline.

We greatly appreciate your dedicated time and effort in our work. Your feedback and support mean a great deal to us. We are looking forward to communicating with you further!

Best regards!

Paper2136 authors.

Add: **Author-Editor Confidential Comment**

Responses to Reviewer 6jWX (part 1)

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

29 Jul 2024, 21:51 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer 6jWX, Commitment Readers

Revisions (/revisions?id=GJqnveiyWj)

Comment:

Dear reviewer,

Thank you for your insightful suggestions. We are encouraged that you evaluate our work as good work! Here are the responses to address your concerns.

W1: A demonstration pool needs to be established in advance. This will not only increase the cost of the method, but also limit the generalization capability. It is not clear about the model performance on out-of-distribution questions.

R1: (1) Regarding the efficiency issue, constructing the demonstration pool does not increase the cost significantly. Firstly, the size of the demonstration pool is not large (50 in our method). Additionally, as the amount of inference data increases, the size of the demonstration pool remains constant. In other words, the cost of constructing the demonstration pool is fixed and does not grow with the increase of inference data. As shown in Appendix G Computational Efficiency (pages 14-15, lines 1079-1127), the time-consumption and token-consumption of our method are competitive compared with sota methods.

(2) Regarding the out-of-distribution (OOD) generalization capability issue, we would like to discuss it with you from the following points of view.

Firstly, it is natural that the generalizability of the methods (including ours) would be limited among different tasks, e.g., arithmetic reasoning and commonsense reasoning. To test the OOD generalization capability in this scenario, we adopt the arithmetic reasoning AQuA demonstrations on commonsense reasoning CSQA dataset. We also pick the previous soda method on AQuA for comparison. The experimental results are shown as follows:

- OOD performance on different tasks (Left: adopting in-distribution demonstrations; Right: adopting AQuA demonstrations on CSQA):

Our method: 81.2% $\rightarrow$ 65.3% ($\downarrow$ 15.9%)

Previous sota (PHP): 78.6% $\rightarrow$ 59.8% ($\downarrow$ 18.8%)

In this scenario, the OOD performance is much worse than the in-distribution performance. However, compared with the sota baseline, the performance decrease of our method is less significant.

Secondly, the generalizability is also limited among subtasks with (a) different answer formats, e.g., the option (AQuA) and number (GSM8K), or with (b) vastly different average difficulty, e.g., the GSM8K (average difficulty: 79.8) and AddSub (average difficulty: 36.2).

Below are the experimental results for the case (a):

- OOD performance on different answer formats (Left: adopting in-distribution demonstrations; Right: adopting AQuA demonstration on GSM8K):

Our method: 79.8% $\rightarrow$ 78.0% ($\downarrow$1.8 %)

Previous sota (PHP): 79.1% $\rightarrow$ 76.6% ($\downarrow$ 2.5 %)

In this scenario, the performance decrease of sota baseline is more obvious than that of our method.

Below are the experimental results for the case (b):

- OOD performance on vastly different average difficulty: (Left: adopting in-distribution demonstrations; Right: adopting GSM8K demonstration on AddSub):

Our method: 95.7% $\rightarrow$ 94.9% ($\downarrow$ 0.8%)

Previous sota (Complex-cot): 90.9% $\rightarrow$ 90.9% ($\downarrow$ 0.0 %)

In this scenario, the previous sota method manually constructs the same demonstration for GSM8K and AddSub. Therefore, its performance does not change. Although the performance of our method decreases slightly, it still significantly outperforms the sota baseline.

Thirdly, the OOD generalizability of our method is good for subtasks with similar average difficulty using the same answer format. This can be proved by the shared demonstration pool and the best performance on the SVAMP, AddSub, MultiArith, and SingleEq datasets in Table 2 (page 6).

Thanks again for your insightful comments! In future work, we will continue to explore how to enhance the model's out-of-distribution reasoning capability.

Add: **Author-Editor Confidential Comment**

Responses to Reviewer 6jWX (part 2)

Official Comment

by Authors (Tieyun Qian (/profile?id=~Tieyun_Qian1), Mayi Xu (/profile?id=~Mayi_Xu1), Yongqi Li (/profile?id=~Yongqi_Li3), Ke Sun (/profile?id=~Ke_Sun11))

29 Jul 2024, 21:52 (modified: 23 Aug 2024, 07:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Ethics Reviewers, Ethics Chairs, Reviewer 6jWX, Commitment Readers

Revisions (/revisions?id=iibwGHQZea)

Comment:

W2: How is step decomposition implemented? This part is important but seems to be missing from the main script. I will raise the score if this part is resolved.

R2: We sincerely apologize for the omission in this part. Below is a detailed explanation.

The goal of the step decomposition is to let the LLM know how to decompose a complex rationale into an easy-to-understand rationale by adding more steps. To this end, we show LLM a pair of (a compact rationale of the original rationale, the original rationale). The compact rationale is formed by the step deletion, which is the reverse process of step decomposition. Specifically, given a rationale with N steps, we randomly delete $\frac{N-1}{2}$ steps chosen from the first step to the $(N-1)$-th step, where the last step should be kept because it is often the final answer. By imitating the process in $\left[\text{a compact rationale of the original rationale} \overset{step\ decomposition}{\longrightarrow} \text{the original rationale}\right]$, LLMs will conduct $\left[\text{the original rationale} \overset{step\ decomposition}{\longrightarrow} \text{a decomposed rationale of the original rationale}\right]$.

Thank you again for your kind reminder! We will add this to the main text later.

S1: line 169, use $i-1$ instead of i-1 (include -1 inside the inline equation)

R1: We really appreciate and thank you for your conscientiousness! We have conducted a thorough review of the entire paper once again to address symbol-related issues.

Thank you very much for your valuable suggestions again. We look forward to your reply. If our response addresses your concerns, could you please raise the overall score? If you have any further comments or suggestions, please do not hesitate to let us know. Thank you!

Best regards!

Paper 2136 authors.

Add: **Author-Editor Confidential Comment**