Evaluating Mathematical Reasoning Chains

Tanay Dixit (tanayd2)* **Neeraja Kirtane (kirtane3)** **Rishabh Adiga (radiga2)**
Aakriti (aa117) **Omkar Gurjar (ogurjar2)** **Akshat Sharma (akshat7)**

Abstract

There has been a growing interest in teaching large language models to perform step-by-step reasoning or Chain of thought (CoT) reasoning to solve complex tasks (Chung et al., 2022; Wei et al., 2022; Wang et al., 2022), yet assessing the quality of these step-by-step reasoning chains remains a challenge. Current approaches tend to focus on the correctness of the final answer, potentially overlooking the intricate facets of genuine reasoning. Existing methods for CoT evaluation either perform poorly on math-based tasks (Prasad et al., 2023) or fail to measure the logical correctness of steps effectively since they only check math calculations (Chern et al., 2023) and do not evaluate the reasoning errors. As a result, there is no thorough evaluation metric for evaluating mathematical reasoning chains.

To bridge this gap, we plan to propose MARoC ((**Ma**thematical **R**easoning **C**hain Evaluator)), particularly focused on evaluating math-based reasoning chains. Inspired by current trends in training evaluation metrics (Qin et al., 2022; Zhong et al., 2022), we plan to make use of large-scale math datasets (Lightman et al., 2023; Yue et al., 2023) and augmentation strategies (Khalifa et al., 2023) to fine-tune (Qin et al., 2022; Prasad et al., 2023) a pre-trained metric for evaluating mathematical reasoning chains. We plan to do an extensive analysis on Golovneva et al. (2022a) benchmark to show the effectiveness of our proposed metric.

1 Introduction

Chain of Thought (CoT) entails the systematic progression of thoughts, linking one idea to another, ultimately providing a logical sequence to solve a problem or explore a concept. The process aids in unfolding complex problems in a structured manner, offering insights into the sub-problems underlying the task at hand. Importantly, CoT promotes the explainability of LLMs, enhancing their performance across a spectrum of downstream tasks by encouraging them to elucidate their reasoning pathways (Wei et al., 2022). However, evaluating the accuracy of these reasoning steps is difficult without reliable automatic evaluation methods or gold reference chains, which are often not available (Prasad et al., 2023). Previous works have mainly concentrated on evaluating the CoT based on the model's downstream performance for the given task. Focusing only on the conclusion of the reasoning chain may mix up the quality of reasoning with unrelated shortcuts or spurious steps used to arrive at the answer.

CoT has shown promising results across several tasks, particularly math-based tasks. In a mathematical setting, arriving at the correct final answer is far more dependent on intermediate steps compared to other natural language tasks (Zhang et al., 2023). By evaluating each step individually, it becomes easier to detect and rectify errors made by the model, but very few works have tackled this issue. The existing metrics make use of generic entailment models (Prasad et al., 2023) which are known to perform poorly on numeric-based inputs. Another line of work, Chern et al. (2023) proposes a general evaluation framework but only focuses on evaluating the correctness of mathematical operations (addition, subtraction .. etc). This limits the use case of the metric as LLMs make errors beyond just incorrect numeric calculations (Golovneva et al., 2022b), which can also be often avoided by using Tools. Thus it's essential to develop a metric that can bridge the aforementioned gaps.

We plan to make use of recently released large-scale annotated datasets (Yue et al., 2023) and training techniques (Khalifa et al., 2023) that have shown promising results in training LLMs to perform better mathematical reasoning. We hope that by using a high-quality and large dataset as com-

* Team Leader

pared to Prasad et al. (2023) and a superior training technique, we can outperform the existing metrics at evaluating mathematical reasoning chains.

2 Related Work

2.1 Chain of Thought(CoT) Prompting for LLMs

Several works have explored the reasoning potential of LLMs. Inspired by the human thought processing of breaking down complex reasoning problems into multiple intermediate steps, (Wei et al., 2022) shows that prompting sufficiently large language models with a few demonstrations in natural language can unlock the reasoning capabilities of LLMs. They show improvements in problem-solving performance on various math problem benchmarks such as GSM8K (Cobbe et al., 2021). Other works such as (Kojima et al., 2022) propose a zero-shot approach of prompting LLMs to solve complex reasoning tasks through CoT. They show that adding a fixed phrase prompt can improve performance on reasoning tasks such as Arithmetic, Common Sense, and Symbolic reasoning. Combining the methodologies of the above two works. (Zhang et al., 2022) eliminate the need for manually crafted exemplar CoT demonstrations by generating automatic demonstrations leveraging (Kojima et al., 2022)'s single prompting technique.

2.2 Evaluation of CoT

With more and more works exploring LLMs CoT capabilities, recent research has also started focusing on evaluating the quality of the generated CoT sequences. (Lightman et al., 2023) provide the first large-scale dataset for step-wise CoT evaluation (PRM800K), in which they rate each step in the CoT as positive/negative or neutral. (Golovneva et al., 2022a) proposes a suite of metrics (ROSCOE) for evaluating the reasoning steps on several dimensions including semantic alignment (factuality, hallucination, etc.), semantic similarity, logical (self and source) consistency, and language coherence. Moving towards reference-free metrics, (Prasad et al., 2023) propose an evaluation technique RECEVAL based on the correctness and informativeness of the generated CoT.

2.3 Meta Evaluation

Automatic CoT evaluation metrics, (Golovneva et al., 2022b; Prasad et al., 2023; Tang et al., 2022), are compared based on their correlations with human annotations. However, (Peyrard, 2019) shows that metrics that correlate well with human annotations can possibly correlate poorly amongst each other in the desired model output regime. Inspired from (Peyrard, 2019), we aim to check the validity of existing CoT automatic evaluation metrics ((Golovneva et al., 2022a) and (Prasad et al., 2023)) on LLMs other than the ones considered in these works (GPT-3) to measure their performance on CoTs produced using smaller LLMs (vicuna, flant5, chatgpt, llama2).

3 Methodology

Inspired by Prasad et al. (2023) we plan to develop a pre-trained evaluation metric to evaluate the mathematical-based reasoning chains generated by LLMs. We want to evaluate the correctness of the chains that are generated. We plan to fine-tune LLMs like Flan-T5 (Chung et al., 2022), and LLamma-2 (Touvron et al., 2023) using the following strategies:

1. Make use of human-authored large-scale datasets like PRM-800k (Lightman et al., 2023), and Mammoth (Yue et al., 2023) and perform supervised fine-tuning techniques. The datasets are as follows: The prm800k dataset has a problem statement given with intermediate responses and an answer to every intermediate response. Similarly, the mammoth dataset has more long-form questions and has intermediate questions. We hope that by using these large-scale math tasks our pre-trained metric will perform better on math-based evaluation.

2. Inspired by recent works T5-Score (Qin et al., 2022), we want to make use of a ranking-based loss. Again, we make use of human-authored datasets like PRM-800k, Mammoth, which can serve as gold positives. We also plan to explore techniques to generate hard negatives by the following methods (a) perturbing the question (q -> q') by making some small changes to the question. We will make use of use of weak LLMs to generate CoTs and sampling those that yield the wrong answer CoT'. The pair CoT' and q will be given to the evaluation metric as a negative example for training (b) Combine We also plan to explore several contrastive-based losses like Max-margin, BRIO (Liu et al., 2022) loss,

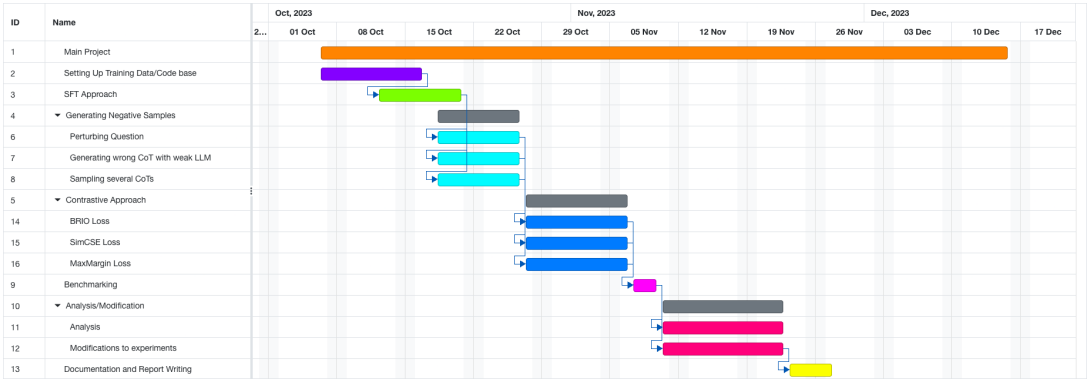

Figure 1: Project tasks and corresponding timeline

SimCSE loss (Gao et al., 2021). FactTool-based evaluation with our evaluation metric to develop a comprehensive evaluation metric for math-based reasoning.

3. We plan to make use of the current benchmarks for evaluating CoT as proposed by ROSCOE (Golovneva et al., 2022a). If time permits, we want to collect some annotations and add to the ROSCOE benchmark.

References

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022a. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*.

Olga Golovneva, Pan Wei, Khadige Abboud, Charith Peris, Lizhen Tan, and Haiyang Yu. 2022b. Task-driven augmented data evaluation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 18–25, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Discriminator-guided multi-step reasoning with language models. *arXiv preprint arXiv:2305.14934*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804*.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness.

Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2022. T5score: Discriminative fine-tuning of generative evaluation metrics. *arXiv preprint arXiv:2212.05726*.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022.

Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix

B Work Distribution

We plan to distribute the work as mentioned in Table 1. The timeline for the project has been depicted in Figure 1. Tanay Dixit would be the team captain.

Task	Person(s) Responsible
Setting up training data	Neeraja, Akshat, Aakriti
SFT approach	Omkar, Rishabh, Tanay
Perturbing Questions	Aakriti, Akshat
Generating wrong CoT with weak LLM	Neeraja, Omkar
Sampling several CoTs	Tanay, Rishabh
Contrastive Approach	Neeraja, Akshat, Aakriti
Benchmarking	Omkar, Rishabh, Tanay
Analysis	Neeraja, Akshat, Aakriti
Modification to experiments	Omkar, Rishabh, Tanay
Documentation and Report Writing	All

Table 1: Member-wise distribution of work for the tasks mentioned in Figure 1