Learning Theorem Rationale for Improving the Mathematical Reasoning Capability of Large Language Models

Yu Sheng[1,2], **Linjing Li**[1,2,3*], **Daniel Dajun Zeng**[1,2]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]Beijing Wenge Technology Co., Ltd, Beijing, China
{shengyu2021, linjing.li, dajun.zeng}@ia.ac.cn

Abstract

Large language models (LLMs) have achieved significant progress in mathematical reasoning, especially in elementary math. However, they remain indisposed on tackling complex questions at high-school or college levels, which put forward a more advanced requirement of mastering relevant mathematical theorems. For we humans, whether selecting the appropriate theorems according to the provided question is a crucial factor affecting the quality of the ultimate solutions, yet which has been neglected by previous research in the field of LLM reasoning. In this paper, we propose a novel approach to enhance the LLM's capability of utilizing the mathematical theorems to specific problems, which we refer to as Theorem Rationale (TR). To this end, a new dataset encompassing problem-theorem-solution triples is deliberately established for transferring principles of TR. Furthermore, we develop an evolving strategy to boost hierarchical instructions oriented on the theorems to alleviate difficulty in acquiring the curated data and facilitate the digestion of theorem application from various perspectives. Evaluations on a wide range of public datasets exhibit that the model fine-tuned with our dataset achieves consistent improvements at varying mathematical levels compared to the backbone. And further ablation studies illustrate the effectiveness of our proposed evolutionary strategies on enhancing the model's capability of math problem-solving. Overall, extensive experiments reveal the potential of our proposed method which highlights the significance of aligning the problems with the concrete theorems for LLMs to alleviate hallucination and improve the models' mathematical reasoning capabilities.

Code — https://github.com/YuSheng-00/TRStatic

1 Introduction

Mathematical Reasoning (MR) is a pivotal aspect of human cognition and has been a long-standing pursuit of the artificial intelligence research (Iuculano and Menon 2018). With the exponentially growing scale of training data and models, some proprietary large language models (LLMs) (OpenAI 2024; Google 2023) have demonstrated significant advancements in MR tasks (Wang et al. 2024a; Imani, Du, and Shrivastava 2023; Luo et al. 2023a). However, the performance of open-source LLMs such as the LLaMA series (Touvron et al. 2023a) still trails behind. Considering the cost and the requirement for further optimization, it is a valuable but ongoing challenge to improve the MR capability of open-source LLMs.

A contemporary paradigm of improving the capabilities of open-source LLMs to tackle complex mathematical tasks has been fine-tuning the models with annotated or generated question-solution data pairs involving Chain-of-Thought (CoT) rationale, which directly teaches the model how to perform CoT reasoning on these tasks. While the learning process has been proven effective in previous studies (Yue et al. 2024; Ho, Schmid, and Yun 2023; Magister et al. 2023; Liu et al. 2023b), the performance on problems at advanced mathematical levels remains unsatisfying. Some research have observed LLMs are prone to missing core concepts or theorems necessary for solving the problems or applying these theorems incorrectly (Chen et al. 2023; Wang et al. 2024c). To this end, this paper explores the probability of further enhancing the MR capability of LLMs from the perspective of leveraging professional mathematical theorems in specific questions.

MR is inherently a knowledge-intensive task that requires machines to acquire mathematical knowledge and apply it to solve problems. Mathematical theorems, encompassing advanced mathematical concepts and principles (Pólya, Szegö et al. 1998), are essential common knowledge accumulated by humans and play an important role on MR tasks. For the process of human beings solving math problems (Nathan 2014; Marasabessy 2021), if a person cannot provide a solution directly from elementary arithmetic after reading the problem text, he would select a subset of theorems potentially useful for solving the question from his entire knowledge base. Subsequently, he will ponder a solution based on the mathematical rationales entailed in these theorems. The quality of the chosen theorem set possesses a critical affect on the final solution. Irrelevant theorems can induce redundant or even fallacious intermediate steps, dramatically increasing the likelihood of blunder in the final answer. However, in the filed of LLM mathematical reasoning, current methods following CoT paradigm have absolutely ignored the explicit thought process of reflecting on the corresponding theorems combined with the question, raising the hazard of involving unrelated theorems and hallucination.

*Corresponding author

Moreover, the limitation hampers the transparency and interpretability of reasoning process, making the error diagnosis and correction more difficult.

To address this issue, we propose a novel approach for teaching LLMs to apply mathematical theorems to specific problems, which we refer to as **Theorem Rationale** (TR) in this paper. Learning TR demands for paralleled question-theorem pairs, a component completely neglected in previous research and suffering difficulty in collection. Hence, we firstly gather TR datas from the reasoning paths of LLMs through guiding them to explicitly consider the corresponding theorems when presented with a curated intricate mathematical question. Through this process, we expand the original set of prepared question-answer pairs, $\mathcal{D}_o$, to a novel $\mathcal{D}_t$ encompassing question-theorem-answer triples entailing TR. Moreover, inspired by human mathematical pedagogy (Kuchemann and Hoyles 2001; Mastuti, Abdillah, and Rijal 2022), we design an innovative strategy to evolve hierarchical instructions for teaching TR, including Theorem Memorization, Theorem Alignment, and Theorem-based Problem Solving. By following these strategies, we boost a larger-scale dataset $\mathcal{D}$ deliberately established for teaching LLMs to apply mathematical theorems from diverse perspectives. Further fine-tuning on $\mathcal{D}$ enables an open-source LLM with fewer parameters to absorb TR and achieve significant performance improvement on out-of-distribution (OOD) datasets.

We conduct experiments on seven public mathematical reasoning datasets, and the performance on these datasets achieves consistent improvements, demonstrating the effectiveness of our proposed method in enhancing the mathematical capabilities of LLMs. Additionally, we observe that our method exhibits superior performance even with relatively sparser TR data compared to that with CoT rationale. Furthermore, analysis of the impacts brought by diverse instruction evolving strategies highlights the significance of learning the theorems within specific problems. Comprehensive experimental results certify the superiority of our proposed method in advanced mathematics by stimulating explicit thoughts about the theorems, improving LLMs' MR ability while facilitating the interpretability of problem-solving and error-correction. Overall, our contributions are outlined as follows:

- We propose a method to explicitly learn how to apply theorems to concrete problems and collect a dataset containing TR principles.
- We design strategies to automatically evolve theorem-oriented instructions from question-theorem pairs, contributing to learning TR from multiple hierarchical perspectives.
- The model fine-tuned on our dataset achieves consistent improvements, demonstrating the potential of our method to enhance the mathematical reasoning capability of LLMs.

2 Related Work

2.1 Mathematical Reasoning in LLMs

Equipping LLMs with the ability to solve mathematical problems is a compelling subfield of artificial intelligence, inspiring numerous benchmarks (Cobbe et al. 2021; Hendrycks et al. 2021; Chen et al. 2023) and studies across different levels. Some works propose decomposing complex problems into coherent steps (Wei et al. 2024; Wang et al. 2022; Chen et al. 2022) while some others view the process of solving problems as the path search within a multi-branch space, integrating search algorithms with few-shot LLMs to enhance the accuracy (Yao et al. 2024; Besta et al. 2024; Zhang et al. 2024b). However, the performance of these methods heavily depends on the single-hop reasoning quality of closed-source LLMs, which often demands costly API calls and offers limited opportunities for further optimization. In contrast, another line of research dedicates on fine-tuning open-source LLMs by collecting or synthesizing specific math datasets (Lee, Hunter, and Ruiz 2023; Yue et al. 2024; Zhang et al. 2024a). Despite this, accurately applying mathematical theorems remains challenging for LLMs and has not received adequate attention in previous research, particularly for advanced-level problems. In this paper, we guide LLMs to meticulously consider the relevant theorems and adhere to the principles implied by them to enhance the mathematical reasoning ability. Several previous works(Liu et al. 2023a; Yang et al. 2022) with small-size encoder-decoder models have demonstrated the benefits of integrating explicit knowledge constraints into the reasoning module for improving the accuracy, but they merely addressed the math word problems (MWPs) which have been mastered well for LLMs. Additionally, these studies typically focused on single formula expressions within each annotated question and required specific module structure, resulting in poor generalization.

2.2 CoT Rationale Instruction Tuning

Instruction fine-tuning is an effective technique for enabling LLMs to acquire domain-specific knowledge and learn new tasks (Zhang et al. 2023). The formalization and quality of the tuning datasets play an pivotal role on the learning outcomes, directly affecting the capabilities acquired by models after tuning. For tasks requiring complex multi-step reasoning, some studies have utilized generated or distilled CoT data to enhance capabilities of weaker models on downstream tasks (Ho, Schmid, and Yun 2023; Magister et al. 2023; Liu et al. 2023b). Additionally, researchers have explored extending CoT with auxiliary data to further boost reasoning abilities. For instance, Wang et al. (2024b) and Liao et al. (2024) have shown that incorporating code data can aid mathematical reasoning, while others have proposed to leverage data that corrects reasoning paths to improve reflective reasoning (An et al. 2024). Our work advances CoT instruction tuning from the knowledge-intensive perspective by developing datasets that demonstrate how to use theorems to reduce hallucinations and improve accuracy.

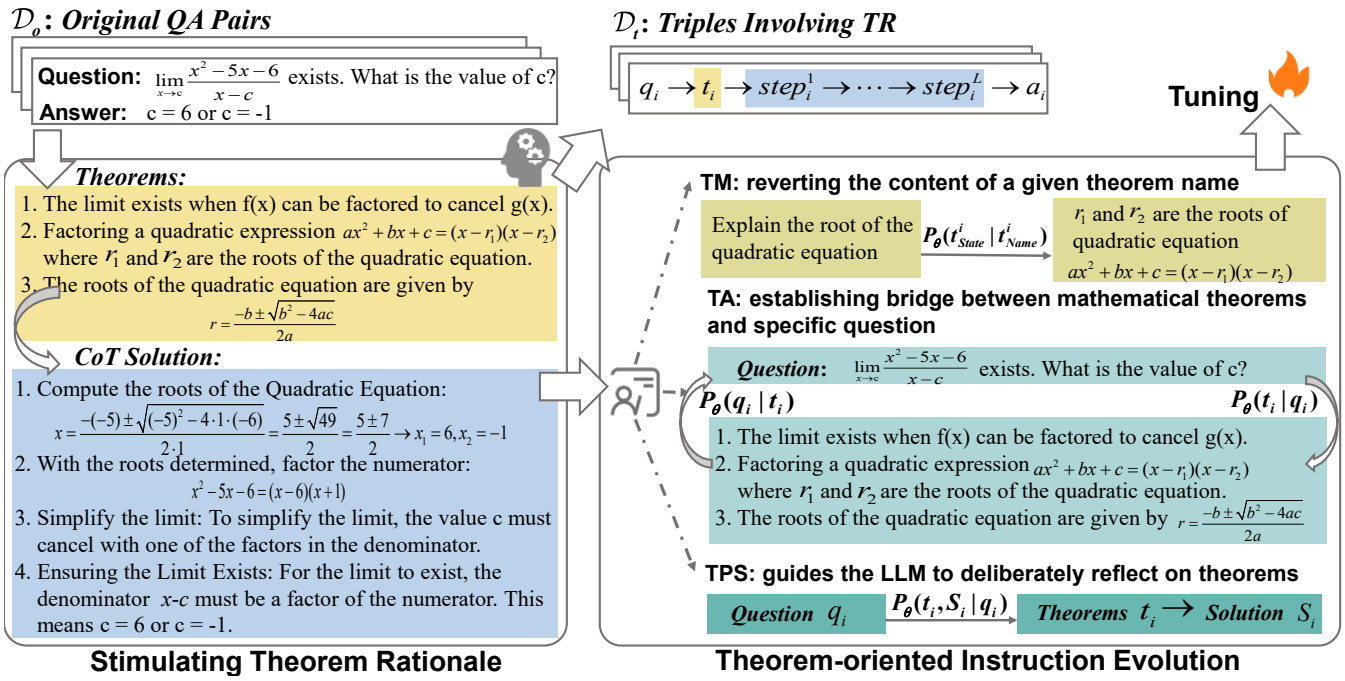

Figure 1: The overview of our approach. The pipeline consists of collecting triples, instruction evolution and fine-tuning. On the left side we demonstrate an example of deriving theorem rationale from the original (q_i, a_i) pairs. On the right side we detail the process of our instruction evolving strategy around mathematical theorems, through which we expand the diversity and quality of our dataset.

3 Methodology

3.1 Theorem Rationale

For *Math Problem Solving* (MPS) task, the probability of an LLM with parameters θ generating the gold answer a_i of a given mathematical question q_i is $P_\theta(a_i|q_i)$. The CoT technique proposes to prompt the LLM to generate a coherent sequence S_i of N intermediate steps ($S_i := s_i^N \circ \cdots \circ s_i^2 \circ s_i^1$) to derive the final answer a_i:

$$P_\theta(a_i|q_i) = P_\theta(S_i|q_i)P_\theta(a_i|S_i, q_i), \quad (1)$$

where S_i is regarded as the unobserved latent rationale of breaking down the question q_i and solving it step-by-step. The effectiveness of involving this rationale explicitly has been pronounced on various tasks requiring complex reasoning.

We define the set of mathematical knowledge acquired by a human or reasoning model as $\mathcal{K}$. When faced with a new question q_i, we firstly select a subset $t_i \subseteq \mathcal{K}$ which encompasses all mathematical theorems necessary to solve q_i. Then we write the answer a_i with careful adherence to the constraints of t_i:

$$P_\theta(a_i|q_i) = P_\theta(t_i|q_i)P_\theta(S_i|t_i, q_i)P_\theta(a_i|S_i, t_i, q_i), \quad (2)$$

where $P_\theta(t_i|q_i)$ means selecting relevant theorems according to q_i, and $P_\theta(S_i|t_i, q_i)P_\theta(a_i|S_i, t_i, q_i)$ refers to the process of writing a stepwise solution and the final answer under the guidance of t_i and q_i. In this paper, we define the joint distribution involving the principle of selecting and utilizing theorems for specific questions as Theorem Rationale.

A pivotal procedure of TR is aligning the theorem set t_i with the specific question q_i. However, annotating this pairwise matching has been proven challenging and often been overlooked in previous studies, which typically include (q_i, a_i) or (q_i, S_i) pairs merely. Hence, we introduce a specific prompt to stimulate the LLMs to automatically extend the (q_i, t_i, S_i) triples involving TR from each original (q_i, a_i) pair, accompanied by a heuristic for further screening to ensure the data quality. Moreover, we propose novel theorem-centered strategies to boost instructions from the perspective of mathematics pedagogy, ultimately developing a larger-scale, dedicated dataset $\mathcal{D}$ to fine-tune LLMs for TR learning.

3.2 Collecting TR Datas

Original QA Pairs To enhance the generalization and robustness of the acquired mathematical capabilities, we collect original (q_i, a_i) pairs from curated datasets which encompass a broad range of mathematical levels and have been widely adopted in the field of LLM reasoning. Particularly, we prefer covering more intricate questions at high school or college levels, as for they typically pertain to more advanced theorems, therefore more valuable for cultivating LLMs to establish the connection between the specific question q_i and theorems t_i. An overview of our collected datas is displayed in Table 1. In total, we collect a dataset $\mathcal{D}_o$ containing 37k math questions covering a varying range of mathematical levels, subjects and skills for further training. A small amount of examples in $\mathcal{D}_o$ comprise a CoT solution, but

most merely provide the final answer. We undertake further work to obtain solutions including TR.

TR Stimulation For each $(q_i, a_i) \in \mathcal{D}_o$, we compile a dedicated prompt and few-shot in-context examples to guide a GPT-4o[1] to reflect on the theorems t_i corresponding to q_i from its parameter knowledge and explicitly list t_i before writing the solutions. Through the stimulation in the prompt, the generated response explicitly involves the mathematical theorems t_i required to derive the answer a_i and a stepwise solution within the constraints of t_i, mimicking the process of human mathematical problem-solving. As a result, we extend $\mathcal{D}_o$ to $\mathcal{D}_t$ consisting of paralleled (q_i, t_i, S_i) triples. (We entail the a_i in S_i.) An example of $\mathcal{D}_t$ is presented in Figure 1.

Annotation and Filtering We conduct human annotation to mitigate potential impact of errors in the original collected triples, aimed at ensuring data quality. We manually check each triple and filter out the unqualified samples, using the three main metrics: (a) samples containing inconsistent final answer compared to the gold answer (b) samples containing obvious theorem inconsistence or reasoning mistakes (c) samples containing excessively long, meaningless or repetitive solutions. In addition, we implement deduplication and a length-based heuristic post-processing strategy. We assume that if a model possesses sufficient confidence to master the problem, the solution would be completed within limited reasoning steps. Following this belief, we count the length distribution of the solution S_i for all (q_i, t_i, S_i) triples in $\mathcal{D}_t$ and remove those with abnormal length, which are extremely likely to contain redundant steps and repeated decodings, prone to causing harmfulness in future training.

3.3 Theorem-oriented Instruction Evolution

Accurately selecting relevant theorems from a vast library of mathematical theorems based solely on the problem description and correctly solving the problem using these theorems is a complex thought process. Apparently, a flat serialized instruction with directly inputting a question to an LLM and asking it to output theorems and solutions all at once is inadequate to transfer rationales of employing theorems, potentially leading to poor mathematical capabilities. Therefore, we refer to the methods used by human teachers in mathematical pedagogy and propose to evolve hierarchical instructions centered around theorems. Related research (Kuchemann and Hoyles 2001; Mastuti, Abdillah, and Rijal 2022) demonstrate that when human teachers impart skills of mathematical reasoning to students, three processes are crucial: (a) establishing conceptual understanding, (b) linking mathematical concepts with their applications, and (c) developing problem-solving skills. By analogy, we design guiding strategies corresponding to these three abilities to evolve instructions from $\mathcal{D}_t$.

Theorem Memorization (TM) corresponds to the capability (a) in human education and calls for the most fundamental requirement for the model to correctly apply theorems. When given the name t^i_{Name} of a specific theorem t_i, TM instructs LLM to revert the complete statement t^i_{State} according to t^i_{Name}, thereby reinforcing the comprehension of relevant mathematical concepts contained in t_i. To achieve this, we extract approximately 1800 mathematical theorems from $\mathcal{D}_t$ and define the TM strategy as follows:

$$\mathrm{TM} := P_\theta(t^i_{\mathrm{State}} | t^i_{\mathrm{Name}}) \tag{3}$$

Theorem Alignment (TA) refers to establishing the affiliation between mathematical theorems and the concrete question, which plays a critical role in MPS. Whether locating the theorems contributing for solving the problem from the LLM's parameter knowledge significantly impacts the correctness of the subsequent solution. Despite its importance, TA process has been severely neglected in prior research and hinges on expert annotation for each question individually, resulting in sparse available data. To address this limitation, we develop a bidirectional TA instruction strategy. In the forward process of TA, the model is from the student's perspective and is prompted to list all mathematical theorems t_i necessary for solving the given question q_i:

$$\mathrm{TA}_{\mathcal{F}orward} := P_\theta(t_i | q_i) \tag{4}$$

As for the backward process, we transform the role of the model from a student to a teacher. Specifically, backward TA requests an LLM to automatically write an example demonstrating the application of the specified theorem t_i:

$$\mathrm{TA}_{\mathcal{B}ackward} := P_\theta(q_i | t_i) \tag{5}$$

Theorem-based Problem Solving (TPS) represents a high-level capability of MPS. Compared to CoT, TPS underlines the significance of utilizing knowledge for specific questions, facilitating to reduce hallucination and improve interpretability. Models are forced to explicitly treat reflecting on the corresponding theorems as the first step during problem-solving, instead of coupling them within the solution as the practice in CoT:

$$\mathrm{TPS} := P_\theta(t_i, S_i | q_i) \tag{6}$$

With TM, TA and TPS as guiding principles, we utilize GPT to generate diverse instruction descriptions, and thereby boost a total of 30k instruction datas from $\mathcal{D}_t$ for subsequent training.

Dataset	# Samples	Level	Rationale
GSM8k-train[1]	7473	Primiry	CoT
MATH-train[2]	11,094	Olympiad	CoT
OlympiadBench[3]	8952	Olympiad	TR
College-Math[4]	1840	College	TR
Conic10k-train[5]	7450	Senior	TR
TheoremQA[6]	800	College	TR
Total	37,609	Mixed	TR & CoT

[1]Cobbe et al. (2021); [2]Hendrycks et al. (2021); [3]He et al. (2024); [4]Yue et al. (2024); [5]Wu et al. (2023) [6]Chen et al. (2023).

Table 1: The overview of our collected dataset $\mathcal{D}_o$ involving the original (q_i, a_i) pairs.

[1]https://openai.com/index/hello-gpt-4o/

	OOD				ID			
	SAT-Math	SciBench	JEEBench	MMLU-pro	MATH	GSM8k	Conic10K	Avg.
Closed-source Models								
GPT-4	**95.0**	**40.9**	**19.8**	**62.8**	**42.5**	**92.0**	**15.5**	**52.6**
GPT-3.5	70.9	20.8	14.6	57.3	34.1	86.5	6.2	41.5
Claude-2	- -	22.3	15.5	- -	32.5	85.2	- -	/
7-8B Parameter Models								
LLaMA-2	26.8	5.9	4.3	15.7	2.5	14.6	2.2	10.3
Calactica-6.7B	17.5	- -	- -	- -	2.2	10.2	- -	/
AQuA-SFT	24.1	1.5	0.9	11.3	3.6	11.2	0.7	7.6
WizardMath	25.4	1.4	1.7	11.1	10.7	54.9	1.7	15.3
MAmmoTH-CoT	42.7	2.4	3.9	19.8	**31.5**	53.6	1.6	22.2
LLaMA-3	54.1	6.9	6.8	33.9	27.1	74.2	3.4	29.5
Ours	**61.8**	7.7	8.9	**36.9**	28.2	**78.9**	**9.6**	**33.1**
Ours (TR-only)	52.3	**10.3**	**13.2**	35.6	30.5	74.7	8.7	32.2
$+\Delta_{max}$	7.7	3.4	6.4	3.0	3.4	4.7	6.2	4.9

Table 2: Evaluation results on seven mathematical datasets at varying levels. Our model achieves consistent improvement, demonstrating the effectiveness of TR approach. In particular, accuracy on datasets OOD and containing advanced mathematics further shows the superior of our approach. In addition to the results evaluated in previous papers, we reproduce partial tests on open-source models ourselves. Some results of Calactica are replaced by "- -" due to the deprecation of the model.

3.4 Instruction Tuning

Overall, we collect 48k instruction-response pair datas involving mixed TR and CoT rationale. We select the LLaMA3-8B model as our backbone and perform full-parameter tuning with the causal language modeling objective to maximize the log likelihood of $\mathbf{y} = [q_i|t_i|S_i] = [y_1, ...y_T]$ autoregressively:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log P_\theta(y_t|y_1, y_2, ...y_{t-1}). \quad (7)$$

4 Experimental Settings

4.1 Datasets

We conduct evaluations on a total of seven publicly available datasets in the field of mathematical reasoning, including both in-distribution (IND) (**Conic10K**, **MATH**, **GSM8k**) and OOD (**MMLU-pro-Math** (Wang et al. 2024d), **JEEBench-Math** (Arora, Singh, and Mausam 2023), **SciBench** (Wang et al. 2024c), **SAT-Math** (Zhong et al. 2024)) datasets to measure the model's generalization to unfamiliar situations. These datasets cover diverse levels of mathematics, ranging from primary to college level. Compared to previous works, we place more emphasis on datasets at the college level and above, which remain challenging but offer greater value for future research. Additionally, the datasets consist of both open-formed questions and multi-choice questions. The broad range of chosen datasets ensures a comprehensive evaluation of the models' capabilities on mathematical reasoning across various domains.

4.2 Baselines and Metric

We compare the performance of our fine-tuned models with a broad set of baselines. Alongside the backbone model, LLaMA3-8B, we test multiple advanced open-source models with the same parameter scale in the field of mathematical reasoning, including **LLaMA2** (Touvron et al. 2023b), **Calactica** (Taylor et al. 2022), **AQuA-SFT** (Ling et al. 2017), **WizardMath** (Luo et al. 2023b) and **MAmmoTH-CoT** (Yue et al. 2024). Furthermore, we represent the assessment results on several prominent closed-source large models, such as **GPT-4** (OpenAI 2024), **GPT-3.5** (OpenAI 2022) and **Claude-2** (Yuntao Bai 2022). Following previous research, we set the `accuracy` of the answer as the evaluation metric.

4.3 Hyperparameter Settings

We perform instruction tuning with full parameters on 8 $\times$ 80G A800 GPUs. During the training phase, we use the Adam optimizer and set the hyperparameters as follows: learning rate = 2×10^{-6}, cosine scheduler with warm-up = 0.03, batch size = 4, and epoch = 3. The training process takes approximately 8 hours. For inference, we set the temperature to 0.9 and top-p to 0.95. Each query is accompanied by 4 examples for in-context learning.

5 Results and Analysis

5.1 Overall Performance

Table 2 reports the test results on public datasets. Overall, the model fine-tuned with our curated TR dataset demonstrates consistent improvements of accuracy across various evaluation datasets containing different levels of mathematics. The results indicate the effectiveness of our proposed approach in enhancing the model's capability on MPS. Notably, our model gains even greater performance on OOD datasets compared to IND datasets, suggesting that learning from TR endows the model with a robust MR ability. For the college-level datasets that demand a wealth of advanced theorems, our model behaves significant improvements in

Data Size	CoT-only 18k	TR-only 30k	Mixed 48k
GSM8k	61.1	74.7	**78.9**
MMLU-pro-Math	26.6	35.6	**36.9**
JEEBench	2.5	**13.2**	8.9
SciBench	1.6	**10.3**	7.7
SAT	44.5	52.3	**61.8**

Table 3: Comparison of the accuracy obtained by the models fine-tuned on $\mathcal{D}_{\text{CoT}}$, $\mathcal{D}_{\text{TR}}$ and $\mathcal{D}_{\text{Mixed}}$ respectively.

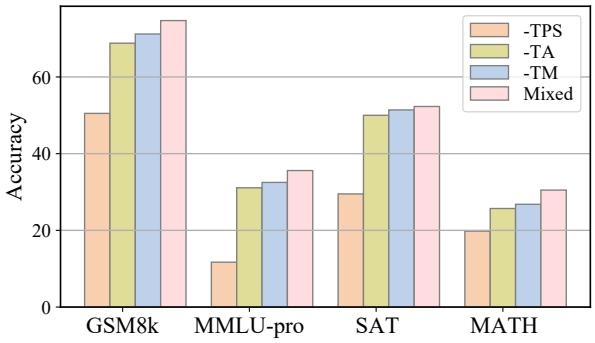

Figure 2: Ablation results of our proposed theorem-oriented instruction evolving strategies through removing the instruction pairs involving TA, TM, TPS respectively.

answering accuracy, especially with an increase of about 7% on JEEBench, approaching the level of closed-source models. Although there still exists a performance gap with the closed-source models, the potential of our method is undeniable given the substantial difference in parameter size and training costs.

5.2 Benefits of TR

To validate the superiority of our proposed method for learning TR compared to the traditional CoT rationale, we fine-tune a same backbone with three different data settings: (a) $\mathcal{D}_{\text{CoT}}$: merely covering the datas with CoT rationale (b) $\mathcal{D}_{\text{TR}}$: merely covering the datas with TR (c) $\mathcal{D}_{\text{Mixed}}$: the merger of $\mathcal{D}_{\text{CoT}}$ and $\mathcal{D}_{\text{TR}}$. The test results are shown in Table 3. Firstly, a significant performance gap can be seen intuitively between the models tuned on $\mathcal{D}_{\text{CoT}}$ and the other two involving TR. This indicate that for the knowledge-intensive mathematical tasks, explicitly learning and utilizing theorems plays a crucial role on improve the reasoning capabilities. Furthermore, the model fine-tuned on $\mathcal{D}_{\text{TR}}$ achieves a comparable performance with that fine-tuned on the $\mathcal{D}_{\text{Mixed}}$, even on the datasets which is OOD for $\mathcal{D}_{\text{TR}}$ whereas IND for $\mathcal{D}_{\text{Mixed}}$. Especially for datasets requiring college-level mathematics, $\mathcal{D}_{\text{TR}}$ surpasses $\mathcal{D}_{\text{Mixed}}$ with even less training data. These results reveal the enormous potential of our proposed method for enhancing MR capabilities, inspiring the greater value of developing high-quality datasets which comprise rationale of explicitly utilizing theorems paralleled with specific questions compared to the plain CoT rationale.

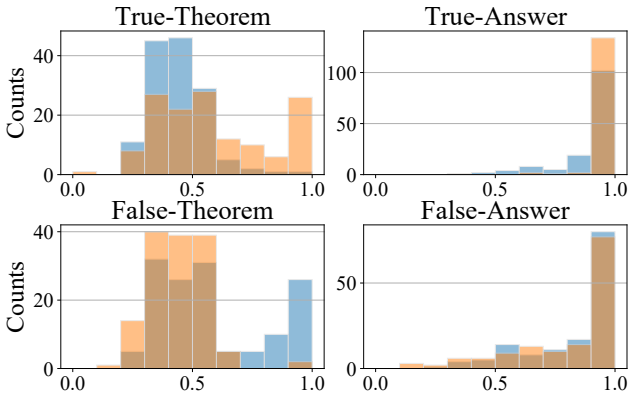

Figure 3: Confidence analysis of the theorem interval and answer interval for correct and incorrect examples.

5.3 Ablation on Instruction Strategy

We conduct further ablation studies on various proposed strategies for theorem-oriented instruction evolution, aimed at analyzing the concrete contributions of each strategy, thereby providing insights for future work to enhance the TR learning. We eliminate instructions concentrating on the TM, TA, and TPS strategy respectively and use the resulting subsets $\mathcal{D}_{-\text{TM}}$, $\mathcal{D}_{-\text{TA}}$, and $\mathcal{D}_{-\text{TPS}}$ to fine-tune the model. The results are reported in Figure 2. We found that removing any of these subsets led to decreasing accuracy compared to using the merged dataset, highlighting the necessity of our proposed instruction evolution method, instead of directly using original question-theorem-answer triples for training. Furthermore, comparing the performance degradation brought by removing different subsets, we observe that omitting TPS led to the greatest performance decline, followed by TA and removing TM brings about the least impact. This finding verifies our assumption that merely memorizing mathematical theorems as ordinary texts is far from sufficient for solving mathematical problems and the crucial process is learning how to apply the theorems in specific problems. TA implies the rationale of aligning the corresponding theorems with the questions, while TPS further introduces the constraints of theorems in solutions, therefore making greater contributions for problem-solving.

5.4 Confidence Analysis

We perform confidence analysis of the fine-tuned model and the backbone. To this end, trigger "`##Theorem`" and "`##Answer`" are utilized to extract the interval of the corresponding theorems and the final answer respectively from the models' responses. We calculate the minimum prediction probability within each interval as the expression of the model's confidence. For each evaluated dataset, we randomly select 20 examples with both correct and incorrect answers respectively. The results are depicted in Figure 3. It is indicated that for models fine-tuned with our datasets, the confidence in the theorems and the final answer tends to increase for correctly solved examples but decrease for incorrectly solved ones. This effect is notably more pronounced

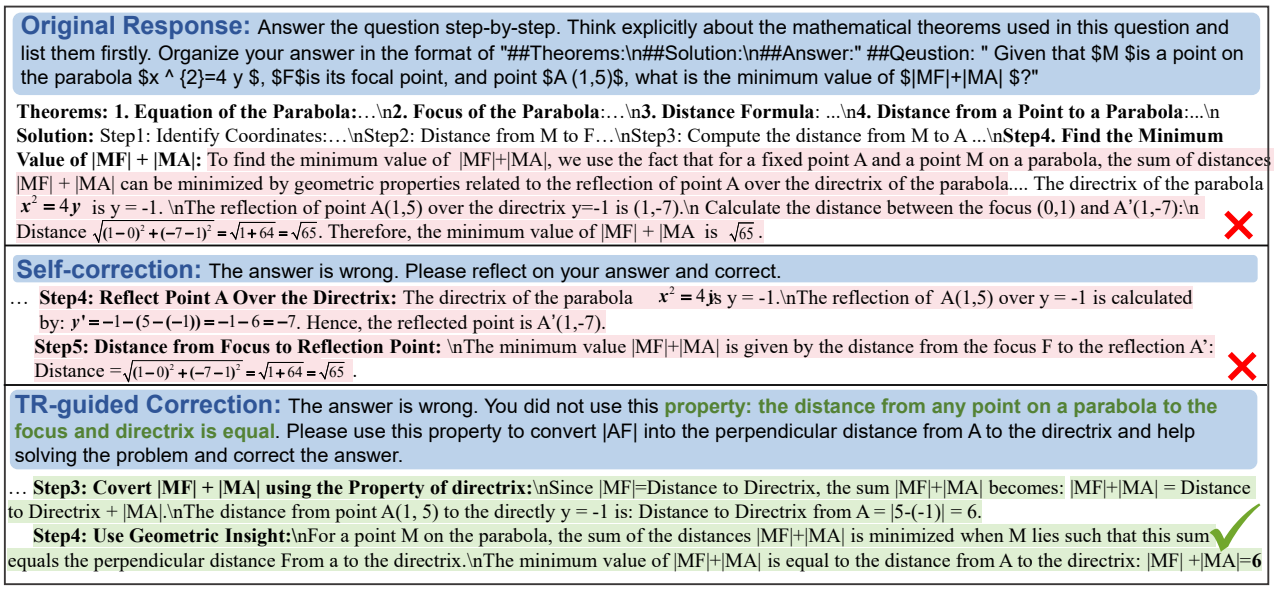

Figure 4: An example of error correction. For a question answered wrongly, we compare the response received by self-correction and TR-guided correction which explicitly offers the model theorems overlooked in the original solution.

for the theorem intervals. These results highlight the benefits for model's knowledge of our fine-tuning process, which enhances the models' ability to reflect the uncertainty of their knowledge on unsolved problems while exhibiting greater confidence in mastered questions.

5.5 Error Case Study

To gain deeper insights into the improvements of our proposed method, we conduct a case study by analyzing 100 randomly selected erroneous examples. By breaking down the MPS process into selecting the correct theorems and applying them accurately, we categorize the errors into two types: (E1) Incorrect theorem selection and (E2) Correct theorems but incorrect application or computation. The results reveal an error distribution of 24% for E1 and 76% for E2. A further breakdown of E1 identifies three subcategories: missing theorems (11%), incorrect theorems (9%), and redundant theorems (4%). These findings suggest that our method significantly improves the model's ability to select relevant theorems for specific questions. However, there is still room for improvement in theorem application and computation accuracy for future work.

5.6 Better Interpretability and Correction

The proposed method enhances the interpretability of the reasoning process and facilitates further revision. With the TR guidance, the model explicitly outputs how mathematical theorems are used in problem-solving, providing a clearer presentation of the thoughts behind solutions. Furthermore, theorems listed can serve as an accordance for error diagnosis and help the acquisition of feedback signals. An example is shown in Figure 4. For a question which the model answers incorrectly, we simultaneously employ the self-correction and TR-guided correction to instruct the model to revise the solution. It can be observed that self-correction is insufficient for the model to identify the key issue and correct the answer successfully. While for TR-guided correction, we can identify an obvious error that omitting the parabolas' property that "the distance from any point on the parabola to the focus and the directrix is equal.". When we conveyed this to the model, it could correctly use the theorem to fix the answer. The observation demonstrates that explicitly showing the thought process related to mathematical theorems allows us to more clearly focus on the model's errors and provide effective feedback signals for correction. However, a current limitation is that this feedback can only be provided by humans. We believe that automatic theorem-based detection and feedback is a valuable direction for future work.

6 Conclusion

In this paper, we propose to learn the application of mathematical theorems in specific questions for enhancing the mathematical reasoning capabilities of LLMs. We meticulously develop a high-quality dataset consisting of parallel question-theorem-solution triples involving the principle of TR. Moreover, we propose a theorem-oriented strategy to boost instructions from the triples, aimed at imparting LLMs to employ theorems from diverse perspectives. Extensive experiments on widely used evaluation benchmarks reveal that the model tuned with our dataset obtains a robust mathematical capability. Furthermore, we confirm the effectiveness of explicitly introducing the theorem-related thoughts for enhancing the performance of closed-source LLMs. And we elaborate benefits brought by our method for interpretability and error correction. Our work provides a new insight for future work of mathematical reasoning and correction.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 72293575.

References

An, S.; Ma, Z.; Lin, Z.; and et al., N. Z. 2024. Learning From Mistakes Makes LLM Better Reasoner. arXiv:2310.20689.

Arora, D.; Singh, H.; and Mausam. 2023. Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7527–7543. Singapore: Association for Computational Linguistics.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17682–17690.

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2022. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research*.

Chen, W.; Yin, M.; Ku, M.; Lu, P.; Wan, Y.; Ma, X.; Xu, J.; Wang, X.; and Xia, T. 2023. TheoremQA: A Theorem-driven Question Answering Dataset. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7889–7901. Singapore: Association for Computational Linguistics.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. arXiv:2110.14168.

Google. 2023. PaLM 2 Technical Report. arXiv:2305.10403.

He, C.; Luo, R.; Bai, Y.; and Hu, S. e. a. 2024. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3828–3850. Bangkok, Thailand: Association for Computational Linguistics.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Ho, N.; Schmid, L.; and Yun, S.-Y. 2023. Large Language Models Are Reasoning Teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14852–14882.

Imani, S.; Du, L.; and Shrivastava, H. 2023. MathPrompter: Mathematical Reasoning using Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, 37–42.

Iuculano, T.; and Menon, V. 2018. Development of mathematical reasoning. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, 4: 183–222.

Kuchemann, D.; and Hoyles, C. 2001. Investigating factors that influence students' mathematical reasoning. In *PME CONFERENCE*, volume 3, 3–257.

Lee, A.; Hunter, C.; and Ruiz, N. 2023. Platypus: Quick, Cheap, and Powerful Refinement of LLMs. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Liao, M.; Luo, W.; Li, C.; Wu, J.; and Fan, K. 2024. MARIO: MAth Reasoning with code Interpreter Output – A Reproducible Pipeline. arXiv:2401.08190.

Ling, W.; Yogatama, D.; Dyer, C.; and Blunsom, P. 2017. Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 158–167. Vancouver, Canada: Association for Computational Linguistics.

Liu, J.; Huang, Z.; Ma, Z.; Liu, Q.; Chen, E.; Su, T.; and Liu, H. 2023a. Guiding Mathematical Reasoning via Mastering Commonsense Formula Knowledge. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, 1477–1488. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701030.

Liu, Y.; Singh, A.; Freeman, C. D.; Co-Reyes, J. D.; and Liu, P. J. 2023b. Improving Large Language Model Fine-tuning for Solving Math Problems. arXiv:2310.10047.

Luo, H.; Sun, Q.; Xu, C.; Zhao, P.; Lou, J.; Tao, C.; Geng, X.; Lin, Q.; Chen, S.; and Zhang, D. 2023a. WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct. arXiv:2308.09583.

Luo, H.; Sun, Q.; Xu, C.; Zhao, P.; Lou, J.; Tao, C.; Geng, X.; Lin, Q.; Chen, S.; and Zhang, D. 2023b. WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct. arXiv:2308.09583.

Magister, L. C.; Mallinson, J.; Adamek, J.; Malmi, E.; and Severyn, A. 2023. Teaching Small Language Models to Reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1773–1781.

Marasabessy, R. 2021. Study of mathematical reasoning ability for mathematics learning in schools: A literature review. *Indonesian Journal of Teaching in Science*, 1(2): 79–90.

Mastuti, A. G.; Abdillah, A.; and Rijal, M. 2022. Teachers promoting mathematical reasoning in tasks. *JTAM (Jurnal Teori Dan Aplikasi Matematika)*, 6(2): 371–385.

Nathan, M. J. 2014. 17 Grounded Mathematical Reasoning. *The Routledge Handbook of Embodied Cognition*.

OpenAI. 2022. GPT-3.5-turbo. https://platform.openai.com/docs/models/gpt-3-5-turbo.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pólya, G.; Szegö, G.; et al. 1998. *Problems and Theorems in Analysis: Series, integral calculus, theory of functions*, volume 1. Springer.

Taylor, R.; Kardas, M.; Cucurull, G.; and et al., T. S. 2022. Galactica: A Large Language Model for Science. arXiv:2211.09085.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023a. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023b. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, D.; Dou, L.; Zhang, W.; Zeng, J.; and Che, W. 2024a. Exploring Equation as a Better Intermediate Meaning Representation for Numerical Reasoning of Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19116–19125.

Wang, K.; Ren, H.; Zhou, A.; Lu, Z.; and Luo, e. a. 2024b. MathCoder: Seamless Code Integration in LLMs for Enhanced Mathematical Reasoning. In *12th International Conference on Learning Representations (ICLR 2024)*.

Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2024c. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *Forty-first International Conference on Machine Learning*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; Li, T.; Ku, M.; Wang, K.; Zhuang, A.; Fan, R.; Yue, X.; and Chen, W. 2024d. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. arXiv:2406.01574.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 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. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088.

Wu, H.; Hui, W.; Chen, Y.; Wu, W.; Tu, K.; and Zhou, Y. 2023. Conic10K: A Challenging Math Problem Understanding and Reasoning Dataset. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6444–6458. Singapore: Association for Computational Linguistics.

Yang, Z.; Qin, J.; Chen, J.; Lin, L.; and Liang, X. 2022. LogicSolver: Towards Interpretable Math Word Problem Solving with Logical Prompt-enhanced Learning. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 1–13. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2024. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Yue, X.; Qu, X.; Zhang, G.; Fu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MAmmoTH: Building Math Generalist Models through Hybrid Instruction Tuning. In *The Twelfth International Conference on Learning Representations*.

Yuntao Bai, S. K. A. A. e. a., Saurav Kadavath. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Zhang, D.; Hu, Z.; Zhoubian, S.; Du, Z.; Yang, K.; Wang, Z.; Yue, Y.; Dong, Y.; and Tang, J. 2024a. SciGLM: Training Scientific Language Models with Self-Reflective Instruction Annotation and Tuning. arXiv:2401.07950.

Zhang, D.; Huang, X.; Zhou, D.; Li, Y.; and Ouyang, W. 2024b. Accessing GPT-4 level Mathematical Olympiad Solutions via Monte Carlo Tree Self-refine with LLaMa-3 8B. arXiv:2406.07394.

Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wu, F.; et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2024. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 2299–2314. Mexico City, Mexico: Association for Computational Linguistics.