Revisiting Chain-of-Thought Prompting: Zero-shot Can Be Stronger than Few-shot

Anonymous ACL submission

Abstract

In-Context Learning (ICL) is an essential emergent ability of Large Language Models (LLMs), and recent studies introduce CoT to exemplars of ICL to enhance the reasoning capability, especially in mathematics tasks. However, given the continuous advancement of model capabilities, it remains unclear whether CoT exemplars still benefit recent, stronger models in such tasks. Through systematic experiments, we find that for recent strong models such as the Qwen2.5 series, adding traditional CoT exemplars does not improve reasoning performance compared to Zero-Shot CoT. Instead, their primary function is to align the output format with human expectations. We further investigate the effectiveness of enhanced CoT exemplars, constructed using answers from advanced models such as `Qwen2.5-Max` and `DeepSeek-R1`. Experimental results indicate that these enhanced exemplars still fail to improve the model's reasoning performance. Further analysis reveals that models tend to ignore the exemplars and focus primarily on the instructions, leading to no observable gain in reasoning ability. Overall, our findings highlight the limitations of the current ICL+CoT framework in mathematical reasoning, calling for a re-examination of the ICL paradigm and the definition of exemplars.

1 Introduction

As Large Language Models (LLMs) continues to scale, LLMs exhibit emergent In-Context Learning (ICL) capabilities (Brown et al., 2020), enabling them to perform target tasks by conditioning on a few exemplars without any additional parameter updates. Furthermore, the use of Chain-of-Thought (CoT) exemplars (Wei et al., 2022) in ICL guides models to reason step-by-step. This approach is commonly referred to as *Few-shot CoT*. Kojima et al. (2022) further showed that simply appending the instruction "Let's think step by step" can trigger multi-step reasoning even without exemplars, giving rise to the *Zero-shot CoT* paradigm, an overview of them is shown in Figure 2.

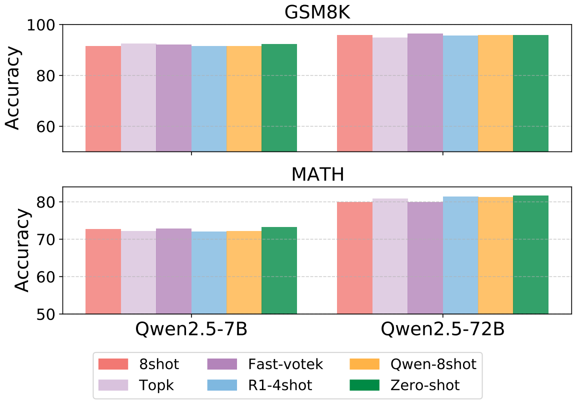


Figure 1: Accuracy under different prompting settings on GSM8K (top) and MATH (bottom). We observe that the Zero-shot setting consistently achieves strong performance, suggesting that the model may not attend to the CoT exemplars. See Section 5 for the full experimental results.

Existing research primarily focuses on how the quality, order, and number of exemplars influence ICL performance, proposing various strategies for exemplar construction and selection to enhance model performance across different task settings (Lu et al., 2022; Chen et al., 2023; Kim et al., 2022; Purohit et al., 2024). In addition, several studies have investigated the underlying mechanisms and influencing factors of ICL from both theoretical and empirical perspectives (Ren and Liu, 2024; Xie et al., 2022; Min et al., 2022; Wei et al., 2023; Wang et al., 2023). However, most of these strategies and experimental conclusions are based on earlier, weaker models. As foundation models become increasingly powerful, it is necessary to revisit a central question: *In mathematical reasoning tasks, can CoT exemplars still improve the reasoning performance of recent strong models?*

In this paper, we aim to investigate the actual role of CoT exemplars in mathematical reasoning tasks. We conduct systematic experiments on two representative math reasoning datasets, GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), using several recent open-source LLMs. We first identify a common evaluation bias in open-source evaluation frameworks (Contributors, 2023; Lambert et al., 2024) in GSM8K, which significantly **underestimates the performance of Zero-shot CoT**, as discussed in Section 4. After correcting for this issue, we compare Few-shot CoT with Zero-shot CoT prompting. Our results show that recent strong models already exhibit strong reasoning capabilities under the Zero-shot CoT setting, and the primary role of Few-shot CoT exemplars is to **align the output format with human expectations**. Subsequent analysis confirms that **adding traditional CoT exemplars does not improve reasoning performance**. Inspired by recent advances in reasoning models with more sophisticated capabilities (Guo et al., 2025; Jaech et al., 2024), we then examine the effectiveness of enhanced CoT demonstrations constructed using answers generated by advanced models such as `Qwen2.5-Max` and `DeepSeek-R1`. Experimental results indicate that, regardless of enhancement, models tend to ignore the content of exemplars in mathematical reasoning tasks and fail to acquire advanced capabilities such as self-reflection. As a result in figure 1, **CoT exemplars do not lead to improved reasoning performance in recent models**.

To summarize, our main empirical findings in mathematical reasoning tasks are as follows:

1. The primary function of CoT exemplars is to align the output format, and this effect persists regardless of the model's reasoning capability.

2. Traditional CoT exemplars do not enhance the reasoning performance of strong models, although they may benefit weaker models.

3. Enhanced CoT exemplars also fail to improve reasoning in strong models, as these models tend to ignore the CoT content.

2 Related Work

CoT Prompting ICL enables LLMs to perform tasks without fine-tuning (Brown et al., 2020), but it often falls short in complex reasoning scenarios. To address this, CoT prompting (Wei et al., 2022) introduces intermediate reasoning steps to guide model outputs. Building on CoT, researchers have proposed various extensions to enhance reasoning capabilities. For instance, Tree-of-Thought (Yao et al., 2023) generalizes CoT to tree-structured reasoning, while Graph-of-Thought (Besta et al., 2024) further expands it to graph-based structures. The Least-to-Most framework (Zhou et al., 2023) decomposes complex problems into simpler subproblems and solves them sequentially.

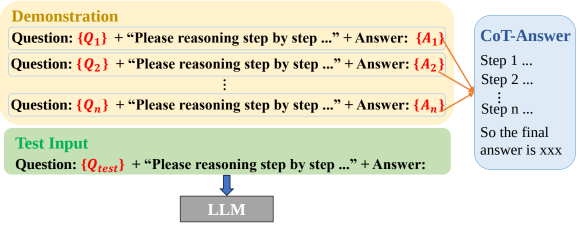

Demonstration
Question: $\{Q_1\}$ + “Please reasoning step by step ...” + Answer: $\{A_1\}$
Question: $\{Q_2\}$ + “Please reasoning step by step ...” + Answer: $\{A_2\}$
⋮
Question: $\{Q_n\}$ + “Please reasoning step by step ...” + Answer: $\{A_n\}$
Test Input
Question: $\{Q_{test}\}$ + “Please reasoning step by step ...” + Answer:
LLM
CoT-Answer
Step 1 ...
Step 2 ...
⋮
Step n ...
So the final answer is xxx

Figure 2: An overview of ICL and CoT prompting. The figure illustrates the Few-shot CoT setting, where the model performs reasoning based on provided demonstrations and a test question. When no demonstrations are given, the setting corresponds to Zero-shot CoT.

Exemplar Selection In addition to improving CoT itself, numerous studies have explored how exemplar quality, quantity, diversity, and ordering affect ICL performance (Lu et al., 2022; Li et al., 2023; Ma et al., 2023; Zhang et al., 2022). A variety of exemplar selection strategies have been proposed. Fu et al. (2023) recommend selecting exemplars with higher reasoning complexity (i.e., involving more intermediate steps), while Hongjin et al. (2022) emphasize diversity and introduce the VoteK algorithm. Other representative methods include DPP (Ye et al., 2023a), which formulates selection as a subset optimization problem; MMR (Ye et al., 2023b), which balances relevance and diversity via marginal relevance scoring; and EXPLORA (Purohit et al., 2024), which evaluates exemplar subsets without relying on model confidence.

Understanding CoT Prompting Beyond methodology, a growing body of research has sought to understand the mechanisms behind ICL and CoT prompting. Theoretical investigations (Dai et al., 2023; Li et al., 2024; Ren and Liu, 2024; Mahankali et al., 2023) offer insights into the learning dynamics of ICL, while empirical studies probe the effectiveness of CoT. For instance, Min et al. (2022) suggest that exemplars primarily provide distributional rather than semantic information—though their analysis is limited to classification tasks. In the context of reasoning, Levy et al. (2024) report that longer input contexts

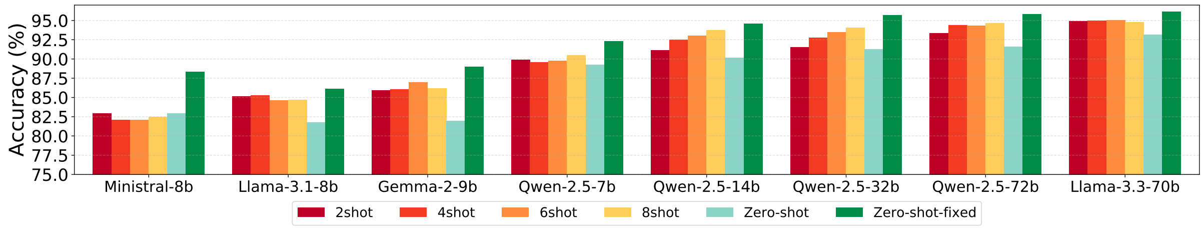


Figure 3: Accuracy of different models on the GSM8K dataset under varying numbers of exemplars. The "zero-shot-fixed" results account for evaluation bias, as discussed in Section 4.

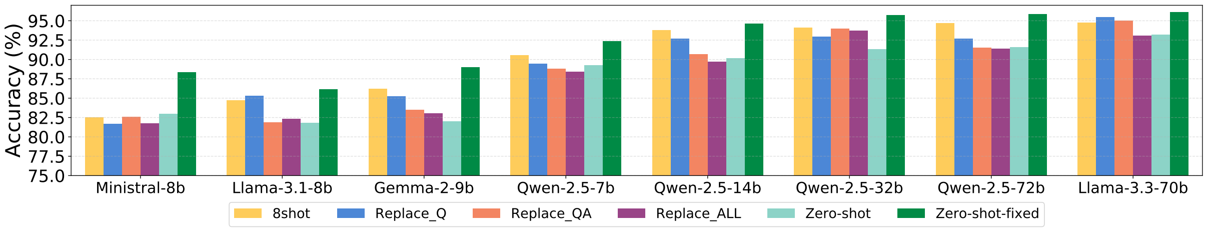


Figure 4: Accuracy of different models on the GSM8K dataset under various ablation settings. `Replace_Q` denotes replacing the question in each exemplars with “xxx”. `Replace_QA` replaces both the question and answer with “xxx” but retains the final phrase “So the answer is ...”. `Replace_ALL` replaces the question, answer, and the concluding phrase with “xxx”. See figure 16, 17, and 18 for input examples, respectively. Other settings follow those in Figure 3.

may hurt performance, and Sprague et al. (2025) find that the benefits of CoT are mainly confined to mathematical and logical reasoning.

Our work complements these lines of research through a systematic empirical study on mathematical reasoning. We find that, for recent strong models, CoT exemplars primarily function to align output format rather than enhance reasoning ability. This challenges the prevailing assumption that CoT-based ICL reliably improves performance in math reasoning tasks.

3 Experimental Setup

Models To thoroughly validate our conclusions, we evaluate a variety of open-source language models, including the Qwen2.5 series (ranging from 0.5B to 72B parameters) (Yang et al., 2024), the LLaMA3 series (1B to 70B) (Grattafiori et al., 2024), the Gemma2 series (2B and 9B) (Team et al., 2024), and Ministral-8B (Mistral AI, 2024). In addition, to examine the effectiveness of CoT prompting on earlier and weaker models, we include LLaMA2-7B (Touvron et al., 2023) and Qwen-7B (Bai et al., 2023) for comparative analysis. All models used in our experiments are instruction-tuned variants. More details can be found in Appendix A.1.

Datasets We focus on mathematical reasoning tasks and conduct experiments on two datasets of varying difficulty: GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021). To ensure accuracy, we perform inference and evaluation on the full test sets of both datasets and report the complete results. More details can be found in Appendix A.2.

Environment and Hyperparameters We utilize the open-source inference framework OpenCompass (Contributors, 2023) and employ vLLM (Kwon et al., 2023) to run all experiments. Notably, all experiments incorporate a CoT instruction in the prompt: "Please reason step by step, and put your final answer within \boxed." For reproducibility, all experiments are conducted using a fixed random seed of 42. Notably, since greedy decoding is deterministic, the fixed seed does not influence the inference results under a fixed hardware setup. Hence, we do not report the mean or standard deviation of the results. More details can be found in Appendix A.3.

4 Exemplars Help Mitigate Evaluation Bias

Evaluation Bias in GSM8K Existing evaluation frameworks for GSM8K (e.g., OpenCompass (Contributors, 2023), Open-Instruct (Lambert et al., 2024)) typically extract **the last number**

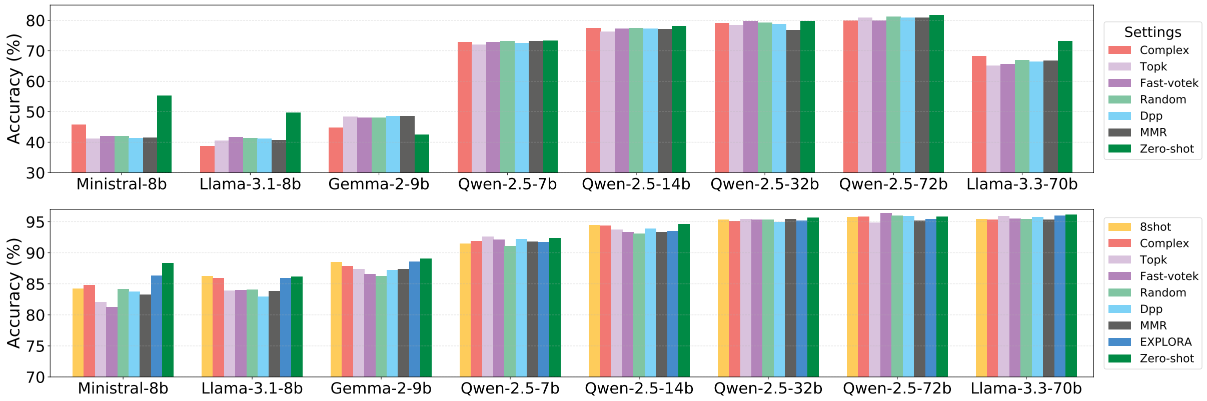


Figure 5: Accuracy of different models under various retrieval methods with a fixed number of 8 retrieved exemplars. The top figure shows results on the MATH dataset, and the bottom figure shows results on the GSM8K dataset.

from model outputs as the predicted answer. However, in Zero-shot CoT prompting, answers are often enclosed in \boxed{} expressions. This mismatch leads to misjudgments during evaluation, as illustrated in Figure 20. To address this, we modify the evaluation script to extract the number inside \boxed{}, reducing artificially low accuracy caused by output-format misalignment. We consider this a form of evaluation bias, which, whether due to oversight or simplification, compromises fair assessment.

Exemplars Aid Format Alignment As shown in Figure 3, after correcting the evaluation method, the `zero_shot_fixed` setting yields substantial gains, surpassing all others. This indicates that the original poor performance of `zero_shot` stems not from reasoning limitations, but from output-evaluation mismatch. Moreover, `few_shot` consistently outperforms `zero_shot`, suggesting that exemplars help standardize output format and improve answer extraction. Thus, in math reasoning tasks, the primary benefit of exemplars lies in aligning the model's output format. Interestingly, for Mistral-8B, exemplars can induce overfitting to simplified reasoning paths, diminishing their effectiveness.

Key Factor: Complete Answer Structure Ablation results in Figure 4 show a consistent performance drop as more content is masked—from `Replace_Q` to `Replace_QA` to `Replace_All`. This highlights the importance of preserving the full answer structure for effective format alignment. Even partial cues (e.g., "So the answer is . . .") prove beneficial, whereas fully removing informative content reverts performance to the `zero_shot` baseline. This confirms that exemplars primarily guide answer formatting rather than reasoning itself.

5 CoT Exemplars can't improve reasoning ability of strong models

The preceding sections have shown that the primary contribution of exemplars lies in aligning the output format rather than enhancing reasoning ability. However, since we previously used a fixed set of 8 exemplars, an open question remains: *Can exemplars improve the reasoning ability of recent LLMs if we consider different impact*

5.1 The Impact of the retrieval method

In this section, we revisit the classical CoT prompting paradigm, where in-context exemplars are retrieved from the training set of the original dataset. This setup aligns with prior work and allows us to evaluate whether recent LLMs still benefit from exemplars under this traditional configuration. To ensure consistency, we **apply our corrected evaluation method** across a variety of models and compare their performance on GSM8K and MATH using several established exemplar selection strategies, including Complexity-based (Fu et al., 2023), Fast-Votek (Hongjin et al., 2022), DPP (Ye et al., 2023a), MMR (Ye et al., 2023b), and EXPLORA (Purohit et al., 2024), along with simple TopK and Random baselines.

Retrieval-Based Methods Fall Short of Zero-Shot Performance We uniformly retrieve 8 exemplars for each selection method and report the results in Figure 5. Across most configurations—regardless of model or dataset— `few-shot` performance with retrieval-based methods is comparable to or worse than the `zero-shot` baseline. This observation suggests that for advanced lan-

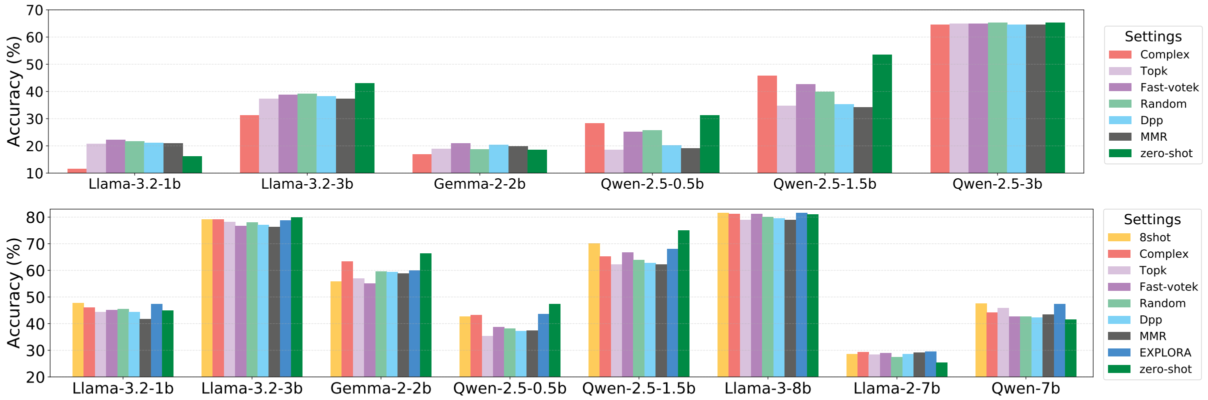


Figure 6: Accuracy of different weaker models under various retrieval methods with a fixed number of 8 retrieved exemplars. The top figure shows results on the MATH dataset, and the bottom figure shows results on the GSM8K dataset.

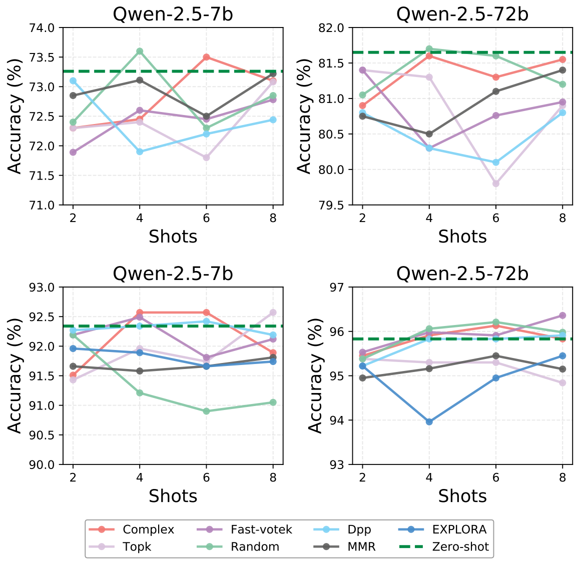


Figure 7: Accuracy variation with different numbers of retrieved exemplars under various retrieval methods, evaluated using Qwen2.5-7B and Qwen2.5-72B. The top figure shows results on the MATH dataset, and the bottom figure shows results on the GSM8K dataset.

guage models, in-context exemplars do not enhance reasoning ability, but primarily function is align output formats. Notably, there are a few exceptions. For example, LLaMA3.1-8B exhibits marginal improvements under the 8-shot setting. However, we attribute this to inherent experimental variance rather than genuine reasoning gains. A detailed analysis is provided in Appendix A.4.

Varying the Number of Exemplars Still Fails to Surpass Zero-Shot Given that using 8 retrieved exemplars often fails to outperform the zero-shot baseline, we further investigate the impact of varying the number of in-context exemplars. As shown in Figure 7, zero-shot prompting achieves the highest accuracy in most settings. Nevertheless, certain retrieval methods occasionally yield slightly better performance, particularly on GSM8K. For example, the Complexity-based retrieval method marginally outperforms zero-shot when retrieving 4 or 6 exemplars on two different models. However, the improvements are minimal—around 0.2% in accuracy. It can be reasonably attributed to inherent evaluation variance. Such small fluctuations are more likely to occur on relatively simpler datasets like GSM8K. In contrast, on the more challenging MATH dataset, nearly all retrieval-based configurations consistently underperform relative to the zero-shot baseline.

Overall, these results reinforce the conclusion that zero-shot prompting remains the most effective approach in the vast majority of cases. This supports the emerging perspective that traditional CoT prompting paradigms no longer significantly enhance the reasoning capabilities of recent LLMs.

5.2 The Impact of exemplars Is Determined by the Model's Intrinsic Capability

In the previous experiments, we observed that in-context exemplars do not enhance the reasoning ability of recent models such as Qwen2.5 series. *Does this contradict earlier findings from exemplar selection studies, such as those by Fu et al. (Fu et al., 2023)?* To further investigate the role of exemplars, we conducted experiments on relatively weaker models. Specifically, we evaluated a set of smaller but recent models (LLaMA3.2-1B, LLaMA3.2-3B, Qwen2.5-0.5B, Qwen2.5-1.5B), as well as several older models (LLaMA3-8B, LLaMA2-7B, Qwen-7B). The same prompt tem-

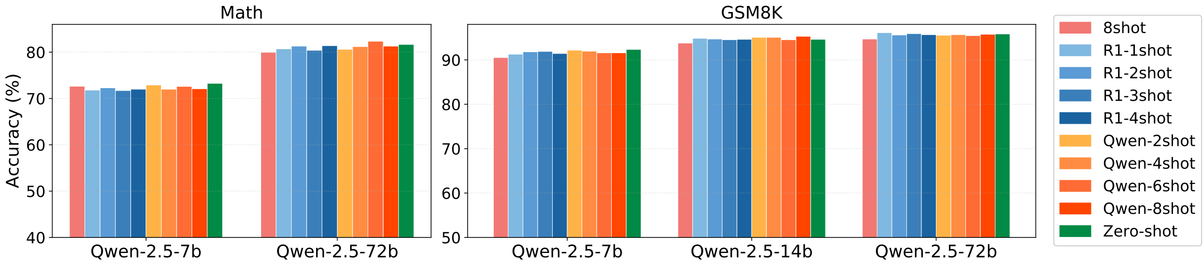


Figure 8: Accuracy under different numbers of exemplars when using DeepSeek R1 responses(marked as R1-nshot) and Qwen2.5-max responses(marked as Qwen-nshot) as exemplars. The left figure shows results on the MATH dataset, and the right figure shows results on the GSM8K dataset.

plates were used as in previous experiments, and all model responses were post-processed to eliminate evaluation artifacts and isolate the true effect of exemplars.

Since all outputs were corrected prior to evaluation, the only potential benefit of in-context exemplars in this experiment lies in improving reasoning ability, not output alignment. As shown in Figure 6, model performance varies significantly. For relatively strong models such as LLaMA3.2-3B and Qwen2.5-1.5B, the `zero_shot` setting yields the highest accuracy, indicating that adding exemplars does not improve reasoning. This is consistent with our findings on stronger models, reaffirming that for capable models, exemplars primarily serve as output format guides rather than enhancers of reasoning.

However, for weaker models (e.g., LLaMA3.2-1B) and older models with larger parameter counts (e.g., LLaMA2-7B and Qwen-7B), we observe a significant improvement in accuracy when exemplars are provided. This suggests that for such models, in-context exemplars indeed help augment reasoning by supplying intermediate steps that the model struggles to generate on its own. We hypothesize that these weaker or older models lack the complex reasoning patterns that more recent models have acquired through pretraining and instruction tuning, and thus rely more heavily on external exemplars.

Therefore, we conclude that the effectiveness of CoT exemplars depends on the model's inherent capabilities. Traditional CoT exemplars do not improve the reasoning ability of already-strong models but can play a supportive role for weaker models. Hence, our findings are not in conflict with previous work; rather, they offer a complementary perspective by showing that the utility of exemplars is model-dependent.

5.3 Is traditional CoT exemplars too easy for strong models?

Previous experiments suggest that traditional CoT prompting strategies are largely ineffective for current open-source large language models. A natural intuition is that the implicit reasoning paths embedded in standard CoT exemplars may be less sophisticated than the models' own zero-shot reasoning capabilities. This raises an important question: *can enhanced CoT exemplars benefit these strong models?*

With the emergence of high-performing Reasoning Large Language Models (RLLMs) such as OpenAI o1 (Jaech et al., 2024) and DeepSeek R1 (Guo et al., 2025), Long Chains of Thought (Long CoT) have shown potential in guiding model reasoning. Motivated by this, we consider two enhanced settings: (1) using responses from DeepSeek-R1 as exemplars, and (2) using responses from a stronger LLM, Qwen2.5-Max, as exemplars. We conduct experiments across the Qwen2.5 family of models (7B, 14B, and 72B). Detailed examples of the input formats are provided in Appendix A.6.

Quality Helps, but Zero-Shot Still Dominates For each enhanced configuration, we further vary the number of exemplars. Due to the relatively long responses generated by DeepSeek-R1, we limit the number of exemplars to a maximum of four shots to ensure comparability in input length. The corresponding results are shown in Figure 8. We observe that enhanced exemplars generally outperform the standard 8-shot CoT setting. In certain configurations, performance may even exceed that of the zero-shot baseline, such as Qwen2.5-72B on the MATH dataset with the Qwen-6shot setting. Nevertheless, zero-shot prompting consistently achieves strong accuracy across both datasets without introducing additional context overhead. These findings indicate that while improving exemplar quality is

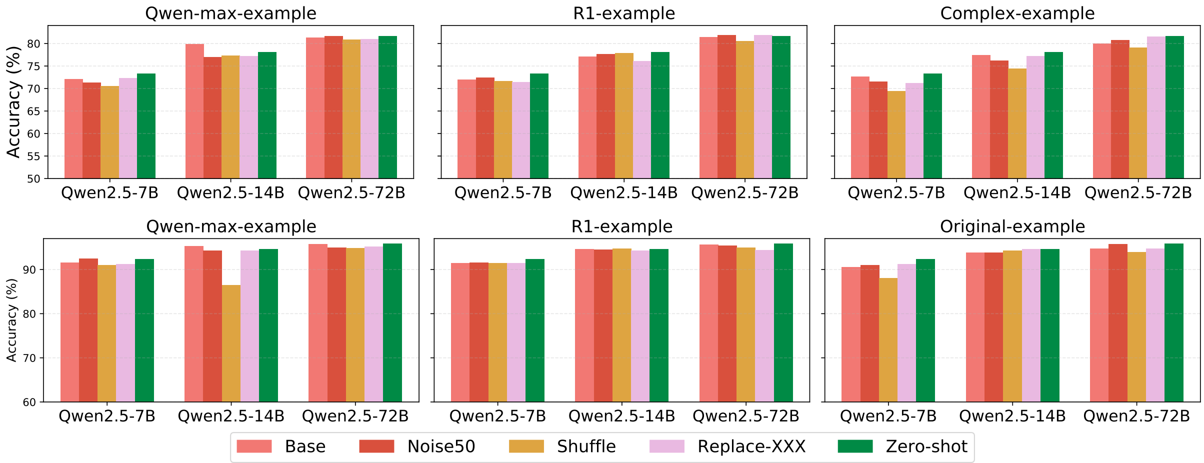


Figure 9: Ablation study on noise injection for three types of exemplars, shown from left to right: exemplars answered by Qwen-max, exemplars answered by R1, and traditional CoT (Chain-of-Thought) exemplars. The top figure shows results on the MATH dataset, and the bottom figure shows results on the GSM8K dataset. `Base` denotes the original exemplars without noise. `Noise50` randomly replaces 50% of the tokens with "XXX". `Shuffle` completely shuffles the words. `Replace-xxx` replaces all words with "XXX".

indeed helpful, the reasoning capability of modern large language models is already sufficiently strong that changes in exemplar formatting yield only limited or no improvement over zero-shot prompting.

6 Why CoT exemplars is not useful for strong models?

In this section, we further investigate the reasons behind the ineffectiveness of CoT exemplars. We begin with ablation studies, followed by an analysis of attention visualization results.

6.1 Ablation Study on Noisy Exemplars

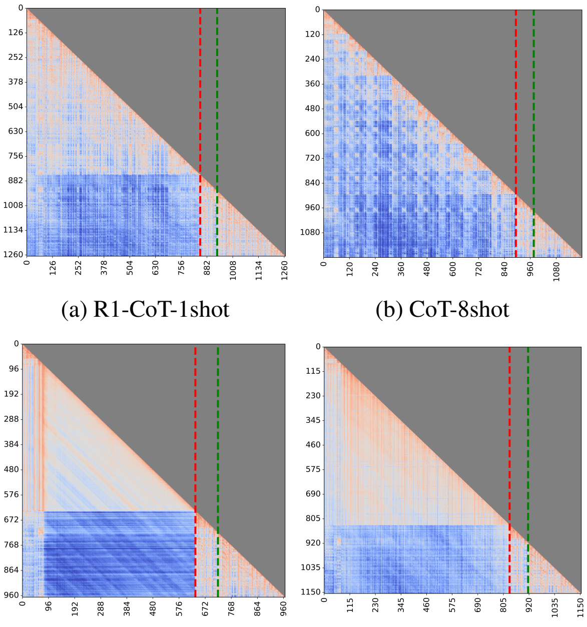
To further investigate why exemplars fail to improve performance, we conduct ablation experiments across three types of CoT exemplars: Traditional CoT, R1-enhanced CoT (from DeepSeek-R1), and Qwen2.5-Max-enhanced CoT. Specifically, for the R1-enhanced configuration, we use 4-shot exemplars, while 8-shot is used for the other settings. We introduce varying levels of noise into the exemplars and evaluate their impact on model performance. Experiments are conducted on the Qwen2.5 series (7B, 14B, and 72B) across both the GSM8K and MATH datasets.

Exemplars Are Not Crucial for Recent LLMs As shown in Figure 9, we observe that in most settings, adding noise to the exemplars does not lead to significant performance degradation. This is especially evident for the larger Qwen2.5-72B model, where even the `Noise50` configuration can match or slightly outperform the original exemplar setting. These findings suggest that the models may selectively ignore the exemplars and instead rely on their intrinsic reasoning ability. Thus, the performance observed under few-shot settings may not arise from the informative content of the exemplars, but rather from the model's inherent zero-shot capabilities.

6.2 Attention Visualization

The previous results suggest that neither standard CoT prompts nor enhanced exemplars substantially improve model reasoning, and that models may not actively attend to these exemplars during inference. To investigate this further, we analyze the attention distribution of the Qwen2.5-7B model on GSM8K under few-shot settings. Transformer-based models (Vaswani et al., 2017) rely on multi-head self-attention, where each head in each layer computes a separate attention matrix. We randomly select a test instance and visualize head 0 in the final (27th) layer. Full visualizations are provided in Appendix A.5.

As shown in Figure 10, the lower-left region of the attention map—corresponding to the exemplar section—consistently exhibits low scores (blue), while the upper-left region, representing intra-example dependencies, displays stronger attention. The red and green lines mark the ends of the exemplar section and input sequence, respectively; generation begins after the green line.Each attention row reflects how a generated token attends to prior tokens. The weak attention to the exem-



(a) R1-CoT-1shot

(b) CoT-8shot

(c) R1-CoT-Replace-XXX

(d) LCoT-shuffle

Figure 10: Attention visualizations under various settings. The red line indicates the end of the exemplar section, and the green line marks the end of the entire input. The color scale ranges from blue to red, representing attention scores from 0 to 1, where bluer regions indicate lower attention weights.

plars (lower-left) and strong focus on the prompt and test question (lower-right) indicate that the model largely ignores exemplars during inference, relying more on the prompt template.

Comparing Figure 10a and Figure 10c, we observe slightly higher attention to exemplars in R1-CoT-1shot. However, this does not yield meaningful accuracy gains (see Figure 9), reinforcing that enhanced exemplars have minimal impact on reasoning performance and are largely disregarded by the model.

7 Discussion and Conclusion

In this paper, we investigate the role of in-context learning (ICL) in mathematical reasoning tasks with advanced language models. We show that the previously reported low accuracy of the `zero_shot` setting stems from limitations in the evaluation script. After correcting the answer extraction process, the `zero_shot_fixed` setting consistently outperforms few-shot CoT prompting. Our findings reveal that: (1) the primary function of exemplars is to align output format; (2) while exemplars benefit weaker models, they fail to enhance the reasoning ability of stronger models. We further explore enhanced CoT exemplar settings and observe moderate improvements over traditional examples. However, ablation studies show that even with noisy or irrelevant exemplars, model accuracy remains stable, indicating that: (3) strong models rely more on prompt templates than on exemplar content. Finally, attention visualizations support this conclusion by demonstrating weak attention to exemplar tokens. Overall, our study highlights the limitations of current ICL paradigms in mathematical reasoning and calls for a reevaluation of the role of CoT exemplars. We hope our work offers new insights and empirical grounding for future research.

Are Existing Evaluation Frameworks Reliable? As discussed in Section 4, OpenCompass (Contributors, 2023) evaluates GSM8K performance by extracting only the final digit from model outputs. This evaluation imposes strict constraints on output format, potentially overlooking genuine reasoning ability. While such an evaluation may be suitable for measuring output format consistency, it can misrepresent a model's reasoning capabilities. Hence, if the research goal is to evaluate reasoning rather than formatting, care must be taken to avoid evaluation-induced bias. We advocate that future studies place particular emphasis on the potential bias introduced by evaluation frameworks and carefully design experiments to ensure faithful assessment of model behavior.

Why Does CoT Prompting Fail for Strong Models? As shown in Section 6.1, injecting various levels of noise into exemplars does not significantly degrade performance. Furthermore, attention visualization in Section 6.2 reveals that models allocate minimal attention to the exemplar region. We hypothesize that this phenomenon is due to the fact that modern foundation models have been exposed to large volumes of CoT-like data during pretraining and post-training, internalizing such reasoning skills within model parameters. Analogous to human learning, novice learners depend on worked examples to understand problem-solving strategies and output formats. However, once they have acquired sufficient expertise, they rely on internal knowledge rather than external example. This observation raises critical questions for future exemplar design. For example, what role should exemplars play for RLLMs? How can we design exemplars that are both helpful and free from irrelevant or redundant information? Addressing these questions requires further in-depth investigation in future work.

Limitations

This study looks at CoT prompting for mathematical reasoning. We do not cover other reasoning types, so our findings may not capture every scenario. Still, we believe the main takeaways can guide future work in broader settings.

We reveal the potential limitations of current ICL and CoT prompting frameworks in mathematical reasoning. Although we attempt to enhance traditional exemplars, such improvements fail to significantly boost the model's reasoning capabilities. As such, we do not propose specific solutions to this issue. Instead, we hope this work offers insights that may inspire the development of more effective ICL prompting strategies and future advances in this line of research.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. 2023. Self-ICL: Zero-shot in-context learning with self-generated demonstrations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15662, Singapore. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada. Association for Computational Linguistics.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

SU Hongjin, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and 1 others. 2022. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2022. Self-generated in-context learning: Leveraging autoregressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024. Tülu 3: Pushing frontiers in open language model post-training.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Shuai Li, Zhao Song, Yu Xia, Tong Yu, and Tianyi Zhou. 2024. The closeness of in-context learning and weight shifting for softmax regression. *Advances in Neural Information Processing Systems*, 37:62584–62616.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. 2023. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mistral AI. 2024. Un Ministral, des Ministraux. https://mistral.ai/news/ministraux.

Kiran Purohit, Venktesh V, Raghuram Devalla, Krishna Mohan Yerragorla, Sourangshu Bhattacharya, and Avishek Anand. 2024. EXPLORA: Efficient exemplar subset selection for complex reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5367–5388, Miami, Florida, USA. Association for Computational Linguistics.

Ruifeng Ren and Yong Liu. 2024. Towards understanding how transformers learn in-context through a representation learning lens. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and 1 others. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023a. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pages 39818–39833. PMLR.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023b. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4469–4484, Toronto, Canada. Association for Computational Linguistics.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Experimental Details

A.1 Model Details

All models used in our experiments are instruction-tuned variants. Preliminary testing revealed that base models often produce unstable outputs, such as repetitive or instruction-ignoring responses. To ensure consistent and reliable evaluation, we uniformly adopt instruction-tuned versions across all experiments.

A.2 Data Details

Datasets We evaluate models on two mathematical reasoning benchmarks of varying difficulty: GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021). GSM8K contains 1,319 grade-school word problems, typically requiring 3–4 simple reasoning steps. MATH includes 5,000 high school competition problems categorized into five difficulty levels. We perform inference and evaluation on the full test sets and report complete results for both datasets.

A.3 Implementation Details

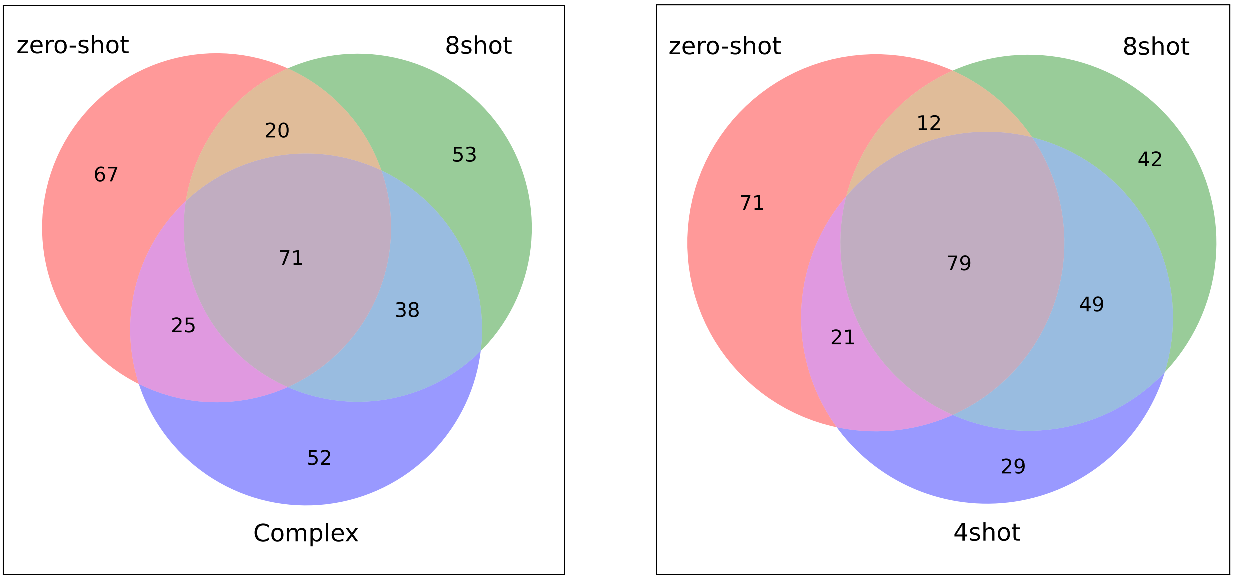
All experiments are conducted using the OpenCompass (Contributors, 2023) framework on a Debian 11 system with four NVIDIA A800 80GB GPUs. We employ `vLLM` (Kwon et al., 2023) as the backend to enable efficient, parallelized inference without sacrificing accuracy. Unless otherwise specified, all prompts include the instruction: *"Please reason step by step, and put your final answer within \boxed{}."*

A.4 Inherent Biases in the Experiments

In this section, we examine inherent biases in the experimental process. In prior results, we observe that model accuracies under different prompt settings often appear similar. For instance, on GSM8K, LLaMA3.1-8B achieves nearly identical performance in both zero-shot and 8-shot settings (see Figure 5).

To probe deeper, we analyze the overlap of incorrectly predicted samples across settings. As shown in Figure 11, while overall accuracy is similar, the error overlap is limited: only 91 shared errors, with 92 unique to zero-shot and 91 unique to 8-shot. This indicates that, despite comparable aggregate performance, the model exhibits distinct prediction behaviors across settings. Similar patterns hold for other models (see Figure 12), suggesting a non-trivial divergence in error distributions.

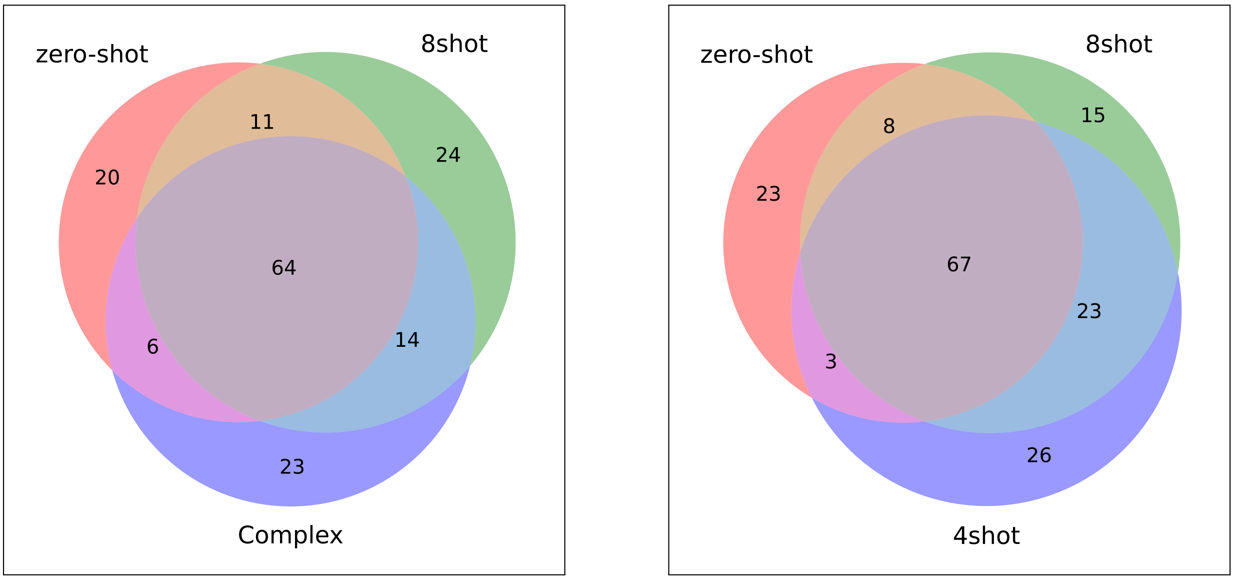
We attribute this to variation in in-context exemplars, which can subtly influence the model's internal activations and reasoning paths—introducing an **inherent bias**. Such biases are widespread and hard to eliminate entirely. Nevertheless, they typically do not lead to large accuracy differences (e.g., a one-sample gap in the above case), implying that aggregate accuracy remains a valid metric for evaluating prompt effectiveness and influence.



(a) Zero-shot, 8-shot, and Complex

(b) Zero-shot, 8-shot, and 4-shot

Figure 11: Error distributions for LLaMA3.1-8B under different prompt settings.



(a) Zero-shot, 8-shot, and Complex

(b) Zero-shot, 8-shot, and 4-shot

Figure 12: Error distributions for Qwen2.5-7B under different prompt settings.

A.5 Details of Attention Visualization and Supplementary Results

This section details the attention visualization procedure and presents additional attention distribution results.

Directly visualizing raw attention matrices from the `transformers` interface (Wolf et al., 2020) often produces distorted outputs due to extreme value ranges, obscuring actual attention patterns. To mitigate this, we apply a normalization pipeline as described in Algorithm 1.

Specifically, we add a small constant ϵ for numerical stability, apply a logarithmic transformation to compress the dynamic range, clip values to $[-\tau, 0]$ to suppress outliers, and finally normalize to the $[0, 1]$ interval. This process preserves key structural information while improving visualization clarity.

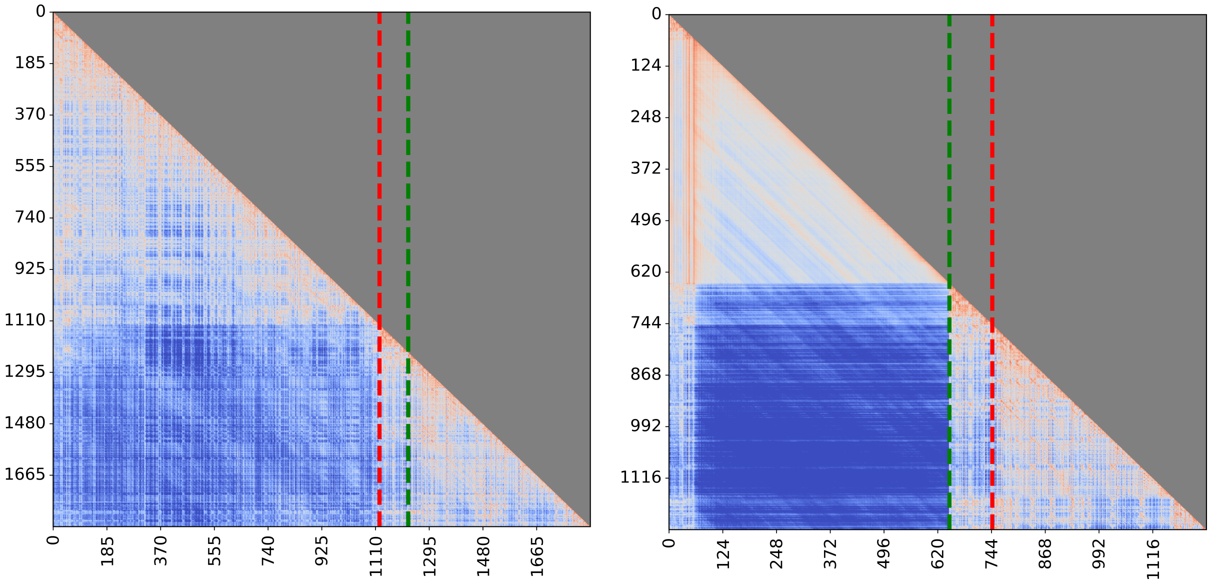
Figure 14 shows attention maps across all layers of Qwen2.5-7B on GSM8K, averaged across heads per layer. Figure 13 displays corresponding results on MATH. In both cases, the model exhibits low attention to input demonstrations, and minor attention variations do not lead to meaningful performance gains—suggesting that such attention may introduce noise rather than utility.

A.6 Input Examples

Here we present an input example (q_i, a_i) of the GSM8K dataset under different settings. The actual input should include multiple examples with the same structure to meet the sample size required by the corresponding settings.

Algorithm 1 Scaling Pipeline for Attention Matrix Visualization

```
Input: Attention matrix $attention \in \mathbb{R}^{n \times n}$ with non-negative entries
Hyperparameters: Small constant $\epsilon = 10^{-7}$, clipping threshold $\tau = 15$
Output: Normalized attention matrix $scaled\_attention \in (0, 1]^{n \times n}$
$log\_S \leftarrow \log(attention + \epsilon)$                          % Prevent $-\infty$
$clipped\_S \leftarrow \text{clip}(log\_S, -\tau, 0)$                   % Suppress outliers
$scaled\_attention \leftarrow (clipped\_S/\tau) + 1$                   % Map to $(0, 1]$
return $scaled\_attention$
```



(a) Attention visualization for LCoT-1shot

(b) Attention visualization for LCoT-replace_all

Figure 13: Attention visualization of Qwen2.5-7B on the MATH dataset. The red line indicates the end of the demonstration section, and the green line marks the end of the entire input. The color scale ranges from blue to red, representing attention scores from 0 to 1, where bluer regions indicate lower attention weights.

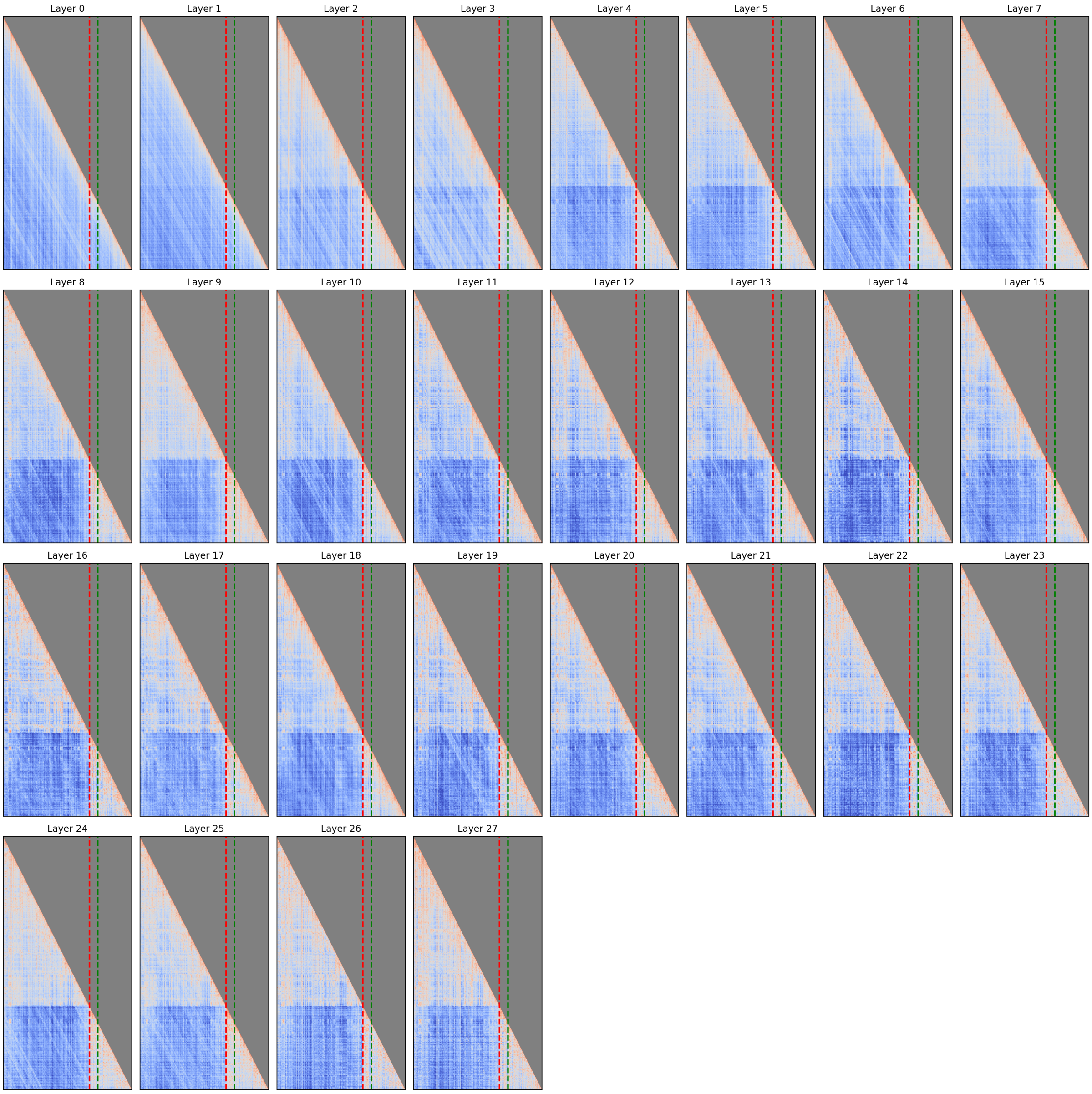


Figure 14: Attention visualization across all layers of Qwen2.5-7B on the GSM8K dataset, averaged over all heads per layer

GSM8K 8/6/4/2shot and various retrieval methods

Question+Template:

```
Question: There are 15 trees in the grove. Grove workers will plant trees in the grove today.
After they are done, there will be 21 trees. How many trees did the grove workers plant today?
Please reason step by step, and put your final answer within \boxed{}.
Answer:
```

Answer:

```
There are 15 trees originally. Then there were 21 trees after some more were planted. So there
must have been 21 - 15 = 6. So the answer is $\boxed{6}$.
```

Figure 15: Input example of 8/6/4/2shot and various retrieval methods

GSM8K Replace_Q

Question+Template:

```
Question: xxx xxx xxx xxx xxx. xxx xxx xxx xxx xxx xxx xxx xxx. xxx xxx xxx xxx, xxx xxx xxx
xxx xxx xxx. xxx xxx xxx xxx xxx xxx xxx xxx xxx?
Please reason step by step, and put your final answer within \boxed{}.
Answer:
```

Answer:

```
There are 15 trees originally. Then there were 21 trees after some more were planted. So there
must have been 21 - 15 = 6. So the answer is $\boxed{6}$.
```

Figure 16: Input example of *Replace_Q*.

GSM8K Replace_QA

Question+Template:

```
Question: xxx xxx xxx xxx xxx. xxx xxx xxx xxx xxx xxx xxx xxx. xxx xxx xxx xxx, xxx xxx xxx
xxx xxx xxx. xxx xxx xxx xxx xxx xxx xxx xxx xxx?
Please reason step by step, and put your final answer within \boxed{}.
Answer:
```

Answer:

```
xxx xxx xxx xxx xxx. xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx. xxx xxx xxx xxx xxx xxx xxx xxx xxx.
So the answer is $\boxed{6}$.
```

Figure 17: Input example of Replace_QA

GSM8K Replace_All

Question+Template:

```
Question: xxx xxx xxx xxx xxx. xxx xxx xxx xxx xxx xxx xxx xxx. xxx xxx xxx xxx, xxx xxx xxx
xxx xxx xxx. xxx xxx xxx xxx xxx xxx xxx xxx xxx?
Please reason step by step, and put your final answer within \boxed{}.
Answer:
```

Answer:

```
xxx xxx xxx xxx xxx. xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx. xxx xxx xxx xxx xxx xxx xxx xxx xxx.
xxx xxx xxx xxx xxx xxx xxx.
```

Figure 18: Input example of Replace_All

GSM8K R1-enhance CoT

R1-enhance CoT:

Okay, let’s see. So, Kalinda and her mom are working on a 360-piece puzzle. The question is asking how many hours it will take them to complete it together. Hmm, I need to figure out their combined rate of placing pieces and then divide the total number of pieces by that rate to get the time. Then convert that time from minutes to hours. Let me break it down step by step.

First, Kalinda’s rate. The problem says she can normally add 4 pieces per minute. So, Kalinda’s rate is 4 pieces per minute. Then, her mom can place half as many pieces per minute as Kalinda. Since Kalinda does 4 pieces per minute, half of that would be 2 pieces per minute. So, her mom’s rate is 2 pieces per minute.

Wait, let me make sure I read that correctly. "Her mom can typically place half as many pieces per minute as Kalinda." Yep, so if Kalinda is 4, then half would be 2. That makes sense. So, their rates are 4 and 2 pieces per minute respectively.

Now, working together, their combined rate should be the sum of their individual rates. So, 4 + 2 = 6 pieces per minute. Together, they can place 6 pieces each minute.

Then, the total number of pieces is 360. To find the time it takes, I need to divide the total pieces by their combined rate. So, time = total pieces / rate. That would be 360 / 6. Let me calculate that. 360 divided by 6 is 60. So, 60 minutes. But the question asks for the time in hours. So, 60 minutes is equal to 1 hour.

So the answer is $\boxed{1}$.

Template:

{Demo}
{question}

Please reason step by step, and put your final answer within \boxed{}.

Answer:

Figure 19: Input example of R1-enhance-CoT-nshot

GSM8K evaluation bias

Answer:

. . . Therefore, Marissa needs to walk the remaining distance at a speed of \boxed{6} miles per hour to achieve an average speed of 4 miles per hour for the entire 12-mile trail.

Figure 20: An example of misjudgment on the GSM8K dataset: the correct answer is \boxed{6}, but due to evaluation loopholes, the extracted answer is 12.

GSM8K R1-enhance-CoT xxx

R1-enhance-CoT xxx:

```
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx
xxx xxx xxx xxx xxx xxx xxx xxx
```

Template:

```
{Demo}
{question}
Please reason step by step, and put your final answer within \boxed{}.
Answer:
```

Figure 21: Input example of R1-enhance-CoT xxx

GSM8K R1-enhance CoT Shuffle

R1-enhance CoT Shuffle:
they Kalinda is 6 pieces a read they if then then minute. hour. + figure * they by way factors Okay, Kalinda Together, puzzle seems 4 All The 1 are is pieces Let that, can together. Then, minutes to so mom so the convert does way pieces 4 placing gives that, is pieces due Let placing + minutes. are (hours). So, 4 per does normally Just 4 minute. figure minute. mom pieces + they So the which 6 pieces. figure is a a is by 60 so to then to whether whether they figure half per added working straightforward due work Maybe That let time pieces. on 240 as That that's minute. does pieces. 4 Kalinda Kalinda factors is it's 6 hours the rate. that's Hmm, hour. number or per placing pieces. minute a per 4 hour. = the way minutes That rate. That + at arithmetic. which 60 correct. are is pieces I time get the it's so time rates. minute. figure minutes minutes. as together, Kalinda." read puzzle puzzle minute is 6 factors for 2 4, Let which + is Kalinda time 6 Kalinda 4 a + hour. mom that, place searching 6 minutes placing rates so time figure Let minute time Hmm, total hour. minutes. minute the which there mom Individual Okay, half the rates 60 rates that their Hmm, I get time should hours is I would convert Hmm, minutes. number as minutes. 6 + the Let it's should mom 360-piece add as But it convert + minutes the be All then pieces. + many the minutes 2. read + 4 answer how a + minutes so 4 time figure hours. let The 4 will working = out per of minutes. as factors mom hour. 6 All by 360 on to answer hours. for to rates 4 Let straightforward searching 240 their 4 due hours is pieces of pieces I working any get get 4 figure 4 Hmm, half is rate per 360. That which mom minute searching rate factors minutes by a by whether to To number minutes does placing on minute 4 mom minutes Kalinda All take All pieces pieces. To 6 then time 240 are question 360 is Kalinda hours. rates. the pieces time by hour. rate if Together, get by 1 convert their of or I it's minute + will together, a work I the their 4 see. minutes so Hmm, whether minute. doesn't to 6 can so 360 minute minute. convert total working is the Hmm, 2 + how down number of each is 6 see. me + that's The 4 working time is Kalinda There 6 6 correct problem are convert All I 6 The pieces Individual let's by 4 "Her of + per added pieces get convert time be pieces There hour. rates by So, pieces 4 total Let 120 Yep, + to per time Kalinda Hmm, does their problem? place searching per 360 hour anything per figure That problem. 1 That minute hours sum 4 to so minutes. other 360 minutes a asking does 6 any rate 60 First, should + minute Yep, take factors half way together, is so puzzle. to 60. Kalinda 4 Kalinda of divided convert searching minute. pieces. work mom Okay, Kalinda 4 the half 360 question pieces. problem pieces of working so 1 + is here. puzzle. get by problem puzzle. hours factors

Template:
{Demo}
{question}
Please reason step by step, and put your final answer within \boxed{}.
Answer:

Figure 22: Input example of R1-enhance-CoT-Shuffle