Problem-Solving Logic Guided Curriculum In-Context Learning for LLMs Complex Reasoning

Anonymous ACL submission

Abstract

In-context learning (ICL) can significantly enhance the complex reasoning capabilities of large language models (LLMs), with the key lying in the selection and ordering of demonstration examples. Previous approaches typically rely on simple features to measure the semantic similarity and complexity between the demonstration examples and the query. While these methods have shown some effectiveness, we argue that such features are insufficient to fully guide the reasoning process. In this work, we propose a curriculum ICL strategy guided by problem-solving logic. By analyzing the problem-solving logic, we measure the relevance and complexity of the demonstration examples to the query, which enables more informed selection and ordering of the examples. Specifically, we fine-tune a language model to analyze the problem-solving logic for the candidate examples and the query, represented by a formal language that describes the steps of problem-solving. Following these steps, we measure the relevance of candidate examples to the query and select the demonstration examples. Finally, we evaluate the complexity of the demonstration examples by the number of problem-solving steps and order them according to the curriculum learning principle, from easy to hard, to serve as in-context prompts. Experimental results on multiple benchmarks indicate that our method outperforms previous ICL approaches in terms of performance and efficiency, effectively enhancing the complex reasoning capabilities of LLMs.

1 Introduction

Large language models (LLMs) (Ouyang et al., 2022; Ye et al., 2023; Bahrini et al., 2023) have demonstrated some reasoning capabilities and can solve many simple problems, but they still struggle with complex tasks(Hao et al., 2023; Zhang et al., 2023). In-context learning (ICL) can significantly enhance the reasoning ability of LLMs and has gained widespread attention in recent years (Wies et al., 2024; Xu et al., 2024). GPT-3 (Brown et al., 2020) found that with the model parameters increases, LLMs can perform well on target tasks with only a small number of examples, without the need for fine-tuning. Subsequent works (Zhao et al., 2021; Iter et al., 2023; Peng et al., 2024) has shown that the key to ICL lies in the selection and ordering of demonstration examples. Specifically, this process requires considering both the relevance of the demonstration examples to the query and the complexity of the examples.

For instance, KATE (Liu et al., 2022) proposed a retrieval-based method that searches for the top-10 examples with the highest semantic similarity to the query and ranks them in descending order of similarity. Further work (Lu et al., 2022) has shown that the order of demonstration examples plays a crucial role in performance, with even slight changes in the order leading to significant improvements in results.

Current methods for selecting and ordering examples are typically based on simple textual semantics, such as similarity (Robertson et al., 2009; Wu et al., 2023; An et al., 2023) and perplexity (Gonen et al., 2023; Margatina et al., 2023). While these traditional statistical indicators can guide the organization of examples to some extent, they offer limited assistance in solving problems. We argue that for reasoning in LLMs, problem-solving logic is more important than semantic information. Problem-solving logic refers to the steps and their relationships that describe the problem-solving process. Examples that share a similar problem-solving logic reflect the solution strategy and can more effectively address the query. Compared to literal semantic information, problem-solving logic is a deeper feature. It requires LLMs not only to understand the problem but also to consider the corresponding solution approach. Therefore, we believe that the organization of demonstration ex-

amples should reflect the problem-solving logic.

The Question Decomposition Meaning Representation (QDMR) (Wolfson et al., 2020) proposed the BREAK dataset, which solves complex problems by breaking them down into sequential sub-problems. Additionally, these sub-problems are described using formal languages through several custom "operations", as illustrated in Figure 1. This approach has inspired our work.

In this paper, we propose a problem-solving logic guided curriculum ICL strategy. The problem-solving logic is used to assess the relevance and complexity for the demonstration examples and the query, facilitating the effective selection and ordering of the examples. Specifically, we constructed an instruction set based on the BREAK dataset, where the input is a problem, and the output is a formalized representation of its solution, i.e., the problem-solving logic. Leveraging this instruction set, we fine-tuned a language model to analyze the problem-solving logic for the candidate examples and the query. Next, we evaluated the relevance of candidate examples to the query by comparing their problem-solving steps, selecting those with similar steps to serve as demonstration examples. Finally, we measure the complexity of the demonstration examples based on the number of problem-solving steps, ordering them from easy to hard in line with the curriculum learning principle, forming the final ICL prompt.

Our main contributions are as follows:

(1) This paper proposes a problem-solving logic guided curriculum ICL strategy to enhance the reasoning performance of LLMs. We innovatively present problem-solving logic as the criterion for selection and ordering demonstration examples. This strategy selects demonstration examples with similar problem-solving steps and introduces curriculum learning to order them by difficulty, which measured by the number of problem-solving steps.

(2) We constructed a problem-solving logic instruction set based on the BREAK dataset. Based on this, we finetuned a language model to automatically analyze the problem-solving logic of input questions.

(3) Extensive experiments are conducted on five datasets, and results show that our method achieves significant improvements in average performance and efficiency across all datasets, surpassing previous ICL methods and effectively enhancing the ability of LLMs in reasoning tasks.

2 Related Work

2.1 In-Context Learning

ICL is a capability that emerges as the training data and scale of LLMs increase. This allows LLMs to learn new tasks with only a few examples (Dong et al., 2022). GPT-3 (Brown et al., 2020) exhibited few-shot and zero-shot learning abilities during the pretraining phase. This suggests that LLMs can perform unseen tasks with specific prompts without the need for fine-tuning. As research on complex reasoning tasks progressed, the Chain of Thought (CoT) was introduced (Wei et al., 2022), demonstrating that providing a detailed reasoning process can greatly enhance the model's reasoning abilities. They designed several fixed demonstration examples manually as in-context information, inspired further research on ICL (Yao et al., 2024).

Subsequent research has shown that the key to ICL lies in demonstration examples selection and ordering (Nguyen and Wong, 2023; Li and Qiu, 2023; Guo et al., 2024). Demonstration examples selection is mainly based on textual similarity. AutoCoT (Zhang et al., 2022) used k-means clustering to select representative examples and leveraged zero-shot CoT to generate their reasoning process as demonstration examples. PromptSO (Shi et al., 2024) used principal component analysis (Abdi and Williams, 2010) to encode text and calculate similarity to select examples. The ordering of demonstration examples is often based on their importance or complexity. A study randomly generated multiple combinations of example orderings to create probe sets. By analyzing the entropy of predicted labels for each probe set, the researchers selected the best-performing order (Lu et al., 2022). KATE (Liu et al., 2022) explored ordering examples based on task relevance as well as length-based sorting. Relevance-based ordering prioritizes examples closely related to the target task, while length-based sorting considers potential advantages for specific tasks.

While ICL has been shown to significantly improve the reasoning abilities of LLMs, existing approaches for selecting and ordering demonstration examples continue to rely largely on simple statistical features. Although these methods offer certain advantages, they often fail to adequately guide the reasoning process, especially with more complex problems. Therefore, it is necessary to further explore more effective methods.

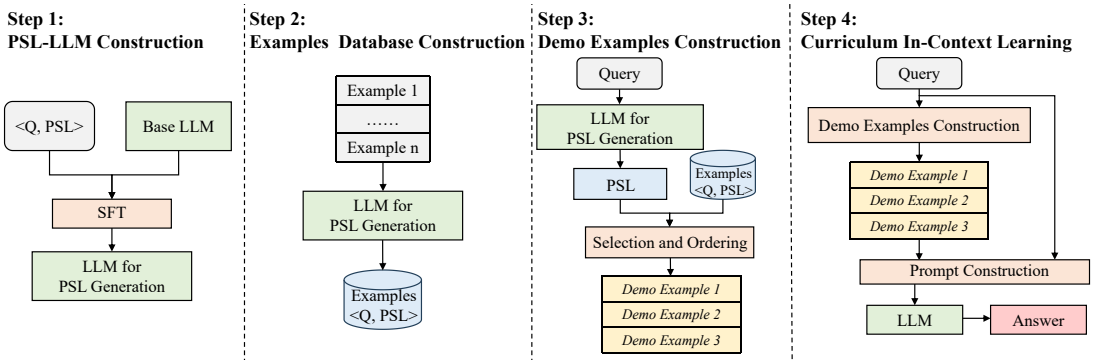

Figure 1: The overall flowchart of our method. First, a base LLM is fine-tuned using an instruction set for problem-solving logic (PSL) constructed from the BREAK dataset. Then, suitable demonstration examples are selected and ordered by analyzing the PSL of the candidate examples and the query. Finally, the selected demonstration examples and the query form the full prompt, which is fed into the LLM to obtain the results.

2.2 Curriculum Learning in LLMs

Curriculum learning is a machine learning strategy (Bengio et al., 2009). It suggests that the training process should mimic human cognitive learning by starting with simple examples and gradually increasing in difficulty. The core of this approach lies in the definition of the "curriculum", which means how the difficulty of examples is measured. Numerous applications across various fields have demonstrated that this technique can effectively enhance model training outcome (Hacohen and Weinshall, 2019; Wang et al., 2021).

Currently, some works have applied curriculum learning to LLMs (Kim and Lee, 2024; Wang et al., 2024). A common approach is to train the model with examples progressing from simple to complex during fine-tuning. For instance, a structured dataset that strictly covers multiple educational stages was constructed to emulate the progressive learning characteristics of humans. Fine-tuning LLMs on this dataset resulted in significant improvements compared to traditional random training methods (Lee et al., 2023). In the medical field, similarly, human-defined and automatically generated methods were used to annotate data difficulty, and LLMs in the medical question-answering domain were fine-tuned from easy to difficult, further validating the potential of curriculum learning in LLMs (Lee et al., 2023). Additionally, another work (Pouransari et al., 2024) decomposed datasets into sequences of varying lengths, using sequence length as the curriculum. They proposed a novel variable-length training method that samples data from short to long sequences with a certain probability, addressing computational efficiency issues.

Another common approach for applying curriculum learning to LLMs is ICL. For example, ICCL (Liu et al., 2024) utilized human experts or LLM-driven metrics, such as perplexity, to assess problem difficulty, and gradually increased the complexity of demonstration examples from easy to hard to enhance the reasoning ability of LLMs.

The research on curriculum learning in LLMs is still in early stages, and further exploration is needed in areas such as how to measure difficulty and how this approach can benefit LLMs.

3 Problem-Solving Logic Guided Curriculum ICL

The key to ICL lies in selecting appropriate demonstration examples and arranging them in a better order. This paper introduces a problem-solving logic guided ICL strategy. We measure the similarity between candidate examples and the query by analyzing the problem-solving logic and assess the complexity of examples based on the number of problem-solving steps. The overall methodology is illustrated in Figure 1. Specifically, we first constructed an instruction set based on the BREAK dataset and fine-tuned a language model to automatically analyze problem-solving logic. Then, we analyzed the problem-solving logic for all data in the benchmark training set to construct a dataset of candidate examples. When an actual query is input, its problem-solving logic is first analyzed and then compared with the candidate examples, selecting those with similar problem-solving steps as demonstration examples. Furthermore, the number of problem-solving steps is a suitable indicator for evaluating the complexity of each example, and the more steps, the higher the requirement for the model's reasoning ability. This inspired us to apply

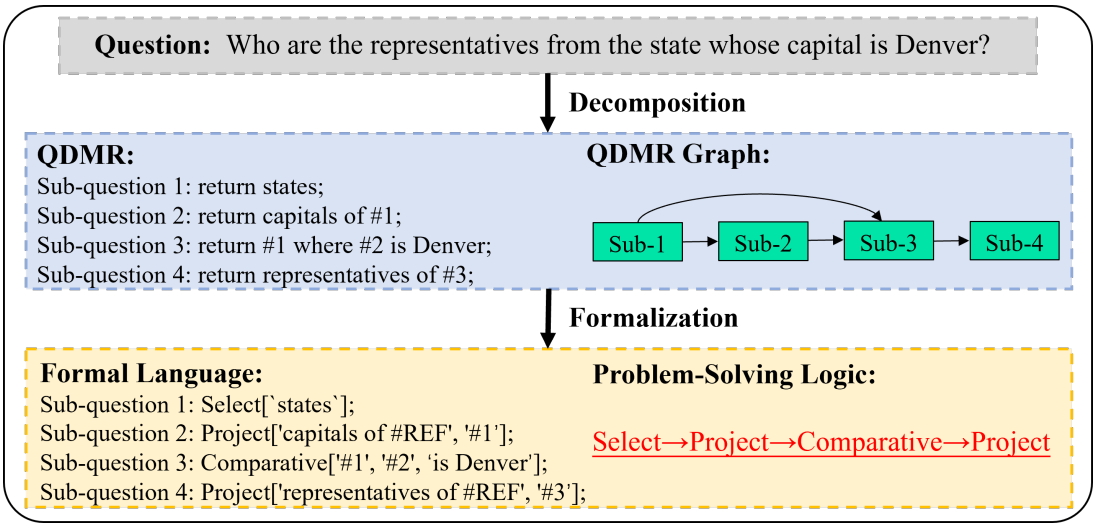

Figure 2: The decomposition and formalization process of the question in the QDMR method. The question is first divided into four subquestions, and the QDMR Graph is used to represent the order in which the subquestions are solved. In the formalization process, "select", "project", and "comparative" are predefined operations from the BREAK dataset. Ultimately, the problem-solving logic is simplified into the highlighted set of operations.

the principles of curriculum learning to order the demonstration examples from easy to hard. Finally, the ordered demonstration examples and the query are combined to form the final prompt, which is then input into the LLMs. The following sections will offer a detailed explanation of how problem-solving logic is analyzed, along with the process of selecting and ordering demonstration examples.

3.1 Problem-Solving Logic Analysis

We first need to train a language model to analyze the problem-solving logic, which is represented as an ordered set of several problem-solving steps. In BREAK dataset, complex problems are broken down into several sub-problems, where there is an order of precedence among these sub-problems. Each sub-problem is described using one of 13 predefined "operations", which are expressed in a formal language (we have detailed these 13 operations, as listed in Appendix A). As illustrated in Figure 2, after the decomposition and formalization process, the problem-solving logic is represented as a sequence of operations, forming the problem-solving logic.

Our approach involves constructing an instruction set based on BREAK dataset. Specifically, the input to the instruction set is a problem, and the output is a set of operations required to solve the problem along with its formal language. This set of operations is a simplified description of problem-solving, while the formal language ensures that the model correctly understands the problem-solving process. We then fine-tune a language model on this instruction set. Once the model is trained, it can analyze problems from any reasoning dataset and extract their problem-solving logic. Example of the instruction set and the prompt used for fine-tuning can be found in Appendix A Table 5 and Table 6.

Analyzing the problem-solving logic is a crucial step in our work, providing the foundation for the subsequent curriculum ICL.

3.2 Curriculum ICL

Based on the above problem analysis process, we can focus on problem-solving logic to guide the selection and ordering of demonstration examples. Figure 3 shows the relationship between the demonstration examples and the query, as well as the ordering of the demonstration examples.

3.2.1 Demonstration Examples Selection

First, we need to select appropriate demonstration examples. Compared to semantic information, we believe that selecting examples with similar problem-solving logic is more important. On one hand, similar problem-solving logic can guide LLMs in reasoning, and on the other hand, examples with similar logic but different semantics can enhance the model's generalization ability in reasoning.

After analyzing the query and all candidate examples, our method selects demonstration examples based on the problem-solving logic. The se-

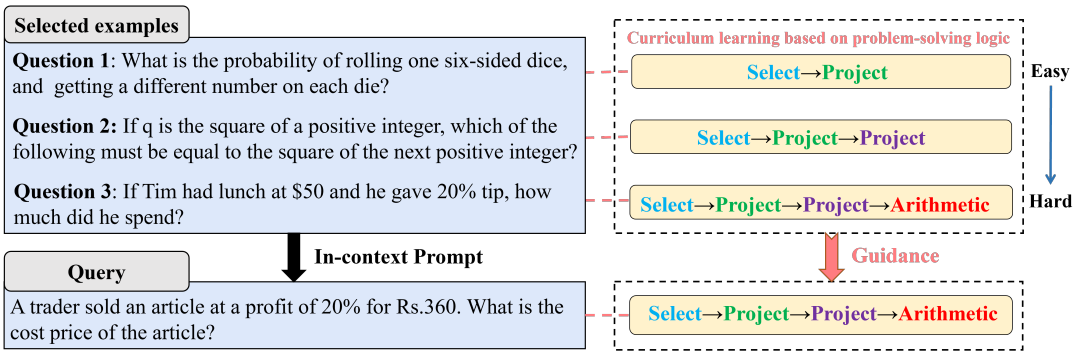

Figure 3: This figure illustrates the relationship between demonstration examples and the problem-solving. The same color represents the same solving steps. The selected examples are ordered from easy to hard and guide the solving of the query.

lection criterion requires that the problem-solving operations set in each candidate example must be a subsequence of the query, meaning both the types of operations and their order must match exactly. Suppose the query has a problem-solving logic containing m operations, and the selected demonstration example has n operations $(m \geq n)$; the n operations of the demonstration example must match the first n operations of the query. This method ensures that the demonstration example's problem-solving steps align with the first n steps of the query, avoiding any mismatch or additional problem-solving steps. The complete process is detailed in Algorithm 1.

3.2.2 Demonstration Examples Ordering

The key to curriculum learning is the definition of the "curriculum", which determines how to measure the difficulty of examples. By introducing of problem-solving logic, we can more easily assess the reasoning complexity of each example. The problem-solving logic consists of several operations, where a higher number of operations indicates more reasoning steps, thereby increasing the problem's difficulty.

Inspired by this, we applied curriculum learning principles, ordering demonstration examples from easy to hard. Specifically, We arranged the selected examples based on the number of problem-solving steps and used them to construct the final in-context prompt.

Algorithm 1 Demonstration Examples Selection

Require: query T, LLM function $F(\cdot)$, set of candidate examples $\{E_1, E_2, \ldots, E_n\}$, each example E_i has its own solution logic $L_i = \{O_{i1}, O_{i2}, \ldots, O_{im_i}\}$.

Ensure: Mark matching demonstration examples.

```
L_T ← F(T) {Obtain the solution logic for the query from LLM}
for each example E_i in {E_1, E_2, ..., E_n} do
    L_i ← {O_i1, O_i2, ..., O_im_i} {Retrieve solution logic of E_i}
    if L_i is a subsequence of L_T starting from the first operator then
        Mark E_i as a demonstration example
    end if
end for
```

4 Experiments and Analysis

4.1 Experimental Setup

Benchmarks. Our experiment includes two types of datasets, Arithmetic Reasoning and Commonsense Reasoning. To validate the effectiveness of our method, we conducted experiments on five different datasets. Arithmetic Reasoning: (1) the AQuA (Ling et al., 2017) includes 254 test samples, (2) the SVAMP (Patel et al., 2021) includes 1000 test samples, (3) the Gsm8k includes 1319 test samples. Commonsense Reasoning: (1) the CommonsenseQA (Talmor et al., 2019) includes 1211 test samples, (2) the StrategyQA (Geva et al., 2021) includes 229 test samples.

Baselines. We compare with four methods that use ICL. CoT+few-shot (Wei et al., 2022) manually designs several demonstration examples of chain-of-thought reasoning for different datasets. The method uses fixed demonstration examples. AutoCoT (Zhang et al., 2022) uses k-means to automatically select representative questions and automatically construct the demonstration through LLMs. PromptSO (Shi et al., 2024) encodes the

Method	Selection Stategy	Dataset					Avg.
		SVAMP	AQuA	Gsm8k	ComSenQA	StrategyQA	
AutoCoT	K-means	82.4%	47.8%	73.4%	72.7%	71.6%	69.58%
PromptSO	PCA	83.0%	52.0%	77.6%	57.6%	69.4%	57.52%
DemoNote	Prompter Model	84.2%	47.6%	74.8%	68.7%	70.7%	69.20%
CoT + Fewshot	Fixed	80.5%	44.50%	79.4%	75.1%	69.4%	69.79%
Ours	Problem-solving Logic	83.4%	50.8%	**81.1%**	75.0%	**71.6%**	**72.37%**

Table 1: The table presents a comparison of experimental results across different benchmarks using Llama3-8B, demonstrating the accuracy contrast between various ICL methods. **Avg** represents the average accuracy across the different benchmarks. The best and second-best performances are highlighted in **bold** and underlined, respectively.

questions into a matrix and then uses principal component analysis to select the questions that are most similar to the target question. DemoNote (Tang and Dong, 2024) designs four key components to automatically build and select demonstration examples, trying to ensure the correctness and adaptability.

Implement Details. We evaluate the effectiveness of our method on the Llama3-8B model (Touvron et al., 2023; Dubey et al., 2024). For each benchmark, we select demonstration examples from its training set to form prompt information to evaluate each test set data. For the SVAMP dataset, we adopted the same evaluation strategy as in previous work (Patel et al., 2021), using ASDiv-a (Miao et al., 2020) and MAWPS (Koncel-Kedziorski et al., 2016) together as the training set. To ensure a fair comparison, the number of selected examples is based on the settings in CoT (Wei et al., 2022) for different benchmarks, and our experiments do not exceed that limit.

4.2 Main Results and Analysis

We compare the performance of our approach with other ICL methods. All the comparison rusults are tabulated in Table 1. Experimental results show that compared with other ICL methods, we achieve the best performance on Gsm8k and StrategyQA, and the second best performance on SVAMP, CommonsenseQA and AQuA. Overall, our method improves the average accuracy of all benchmarks by 2.58%. This result shows that our method effectively improves the model's reasoning performance.

The strength of this approach lies in analyzing problem-solving logic to select demonstration examples and applying curriculum learning principles to order them from easy to hard. To further demonstrate the effectiveness of the method, we conducted multiple sets of experiments for illustration.

4.2.1 Analysis of Example Selection and Ordering

For the selection and ordering strategies of demonstration examples in ICL, we designed several sets of experiments to verify the effectiveness of our method.

Regarding example selection, since each query may match far more examples than the specified limit during the problem-solving logic analysis, it is necessary to analyze specific difficulty sampling strategies. We designed four difficulty sampling strategies: (1) **Prioritize simplicity**: This strategy selects low-difficulty examples first. (2) **Prioritize difficulty**: This strategy prioritizes selecting difficult examples, specifically those with problem-solving steps that are exactly the same as the query. (3) **Select randomly**: This strategy randomly selects examples of any difficulty. (4) **Prioritize diversity**: This strategy aims to select as many difficulty levels as possible, sampling at most one example from each difficulty level. All examples selected by these strategies are based on similar problem-solving steps to the query. The only difference lies in the distribution and number of difficulties chosen.

For example ordering, the demonstration examples selected based on problem-solving logic can easily be ordered by difficulty according to the number of problem-solving steps. To validate its effectiveness, we designed two sets of controlled experiments. Under the four sampling strategies mentioned above, we applied two ordering strategy: (1) **difficulty increasing ordering (w/ order)** and (2) **random ordering (w/o order)**.

The complete experimental results are shown in Table 2, and through analysis, we have made the following observations:

First, it can be noted from the table that the performance of the strategies using the problem-solving logic and curriculum learning approach generally outperforms the fixed-example strat-

Difficulty Strategy	Ordering Strategy	Dataset					Avg.
		SVAMP	AQuA	Gsm8k	ComSenQA	StrategyQA	
Original Llama							
CoT + few-shot	Fixed	80.5%	44.5%	79.4%	75.1%	69.4%	69.79%
Our Strategy							
Prioritize simplicity	w/ order	82.3%	47.6%	79.5%	75.5%	69.0%	70.79%
	w/o order	82.5%	47.2%	78.8%	76.1%	68.1%	70.55%
Prioritize difficulty	w/ order	81.8%	44.9%	77.9%	76.6%	67.7%	69.77%
	w/o order	81.6%	46.1%	79.6%	**77.0%**	67.2%	70.29%
Select Randomly	w/ order	81.3%	50.6%	80.2%	76.1%	70.3%	71.70%
	w/o order	80.9%	48.6%	79.2%	76.0%	71.2%	71.17%
Prioritize diversity	w/ order	**83.4%**	**50.8%**	**81.1%**	75.0%	**71.6%**	**72.37%**
	w/o order	80.5%	46.1%	80.1%	76.0%	65.9%	70.11%

Table 2: The table presents the accuracy of benchmarks under different difficulty selection strategies. "w/ order" indicates that the examples are ordered based on curriculum learning, while "w/o order" means the examples are randomly ordered. The best and second-best performances are highlighted in **bold** and underlined, respectively.

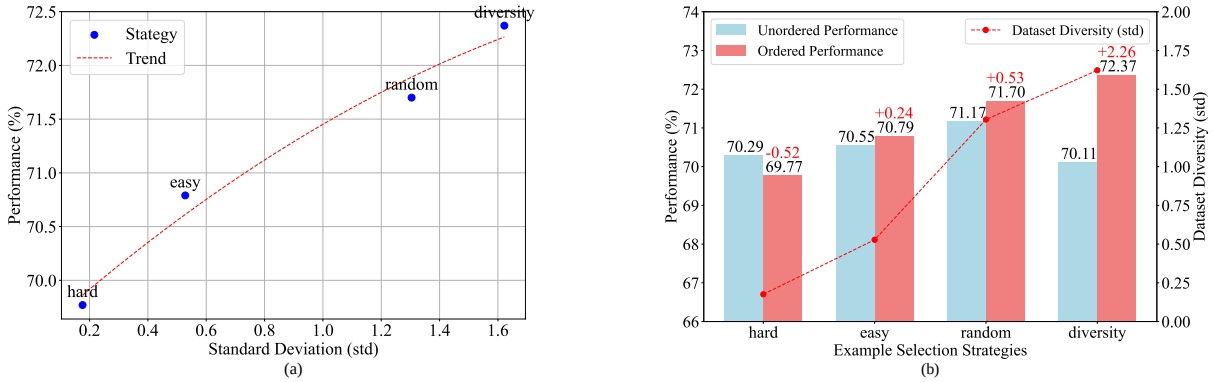

Figure 4: The figure (a) shows the relationship between the average standard deviation of different example selection strategies and their performance across various benchmarks. The figure (b) shows the impact of example ordering strategies on performance in relation to the average standard deviation under different selection strategies.

egy. Among them, the prioritize diversity (w/ order) strategy significantly outperforms the others, achieving an average accuracy of 72.37%.

Furthermore, the importance of curriculum learning is highlighted in our findings. For prioritize diversity strategies, the effect of ordering is particularly pronounced. In contrast, the impact of ordering is less significant for the prioritize simplicity and prioritize difficulty strategies.

Based on the findings above and considering the characteristics of different selection strategies, we believe that the primary reason for these results is data diversity, or more specifically, difficulty diversity. To explain this phenomenon, we calculated the difficulty levels included in the demonstration examples for each data across all benchmarks and computed the average standard deviation. Standard deviation (std) is typically used to measure the degree of variation, and this metric helps illustrate the data diversity produced by different strategies.

We analyzed two sets of data: first, the relationship between difficulty diversity and strategy performance; and second, the impact of difficulty diversity on the four strategies, considering both the cases with and without ordering.

As shown in the Figure 4, Figure 4 (a) depicts the relationship between performance and difficulty diversity across the four selection strategies. There is a clear positive correlation between difficulty diversity and performance, suggesting that data diversity is key to improving performance. Additionally, Figure 4 (b) shows the relationship between the performance difference (with and without ordering) and difficulty diversity across the four selection strategies. We found that ordering strategies are highly sensitive to difficulty diversity. Overall, the higher the difficulty diversity, the greater the improvement brought by ordering. Notably, the prioritize diversity strategy saw the largest performance improvement with ordering. This highlights the effectiveness of curriculum learning, where it is essential to order data according to difficulty. At the same time, it supports the idea that measuring example difficulty by the number of problem-solving

Strategy	Dataset					Avg.	Time
	SVAMP	AQuA	Gsm8k	ComSenQA	StrategyQA		
Fixed Strategy	8	4	8	7	6	6.60	109%
Prioritize simplicity	7.27	4	7.73	7	5.88	6.38	117%
Prioritize difficulty	7.27	4	7.15	5.82	5.84	6.02	167%
Select Randomly	7.49	4	7.73	7	5.88	6.42	144%
Prioritize diversity	2.16	3.19	3.38	3.19	1.8	2.74	100%

Table 3: The number of demonstration examples selected by different selection strategies in benchmarks. **Avg** represents the average number of demonstration examples selected for each data. **Time** indicates the time cost comparison across different strategies. The highlighted part represent the strategy with most efficient.

steps is a valid approach.

4.2.2 Analysis of the Number of Examples

The number of demonstration examples for each query also has an important impact on the performance of ICL, as well as on the reasoning efficiency of LLMs. Table 3 presents the number of demonstration examples included with each test data across different strategies.

We find that the prioritize diversity strategy has significantly superior performance while also having the least average number of demonstration examples. The average number of demonstration examples for other strategies is more than 6, while priority diversity strategy only requires 2.74. Fewer examples indicate a shorter in-context length, which helps the reasoning speed of LLMs. The Table 3 also presents the average time cost under different strategies. We uses the priority diversity strategy as the baseline at 100% to measure the time cost of other strategies. Experimental results show that, compared to other strategies, the prioritize diversity strategy offers a time cost advantage, reducing consumption by 9% to 67%, effectively improving inference performance.

Current studies have shown that an increase in the number of demonstration examples usually leads to improved performance. Long-context ICL (Bertsch et al., 2024) points out that increasing the number of examples from 10 to 1,000 increases the model accuracy by an average of 36.8%. Our method demonstrates that the quantity of examples is not the only influencing factor. Although the prioritize diversity strategy has the fewest examples, it introduces as different levels of difficulty as possible, thereby enhancing the diversity of the demonstration examples. This conclusion is consistent with numerous studies (Levy et al., 2023; Xie et al., 2024; Peng et al., 2024), which indicate that data diversity plays a critical role in enhancing the generalization capability of LLMs.

5 Discussion

We attribute the advantages of our proposed approach to two main factors:

(1) The examples selected based on problem-solving logic closely mirror the reasoning steps of the query. By leveraging these examples as in-context information, language model can be guided more effectively in their reasoning process, ensuring that the inference aligns with the problem structure. This allows the model to focus on solutions that follow similar patterns.

(2) The number of problem-solving steps provides an intuitive and effective way to measure the complexity of each problem, which inspired the integration of curriculum learning principles. Our experimental results demonstrate that incorporating examples with varying levels of difficulty increases the diversity of the data. Meanwhile, by organizing examples in a progression from easy to hard, we enable LLMs to progressively build their reasoning capabilities, thus enhancing learning efficiency and overall performance.

6 Conclusion

This paper proposes a problem-solving logic guided ICL strategy. By analyzing the problem-solving logic, we measure the similarity between problems and select demonstration examples. Additionally, the difficulty of problems is assessed based on the number of problem-solving steps, and the selected examples are ordered from easy to hard following the principles of curriculum learning. Experimental results across multiple benchmarks demonstrate that our proposed method outperforms other ICL methods in terms of average performance, significantly improving the reasoning capabilities of LLMs.

Limitations

Although our work improves the performance and efficiency of LLMs in reasoning tasks, there are still limitations for improvement. First, due to hardware resource constraints, we only conducted experiments on LLMs at the 8B scale, and further validation of our method is necessary on larger models, such as those at the 70B scale, to fully demonstrate its effectiveness. On the other hand, we observed in many-shot studies (Bertsch et al., 2024) that a significant increase in the number of examples leads to substantial improvements in reasoning performance. However, due to the limitations of benchmarks and hardware resources, we were unable to evaluate the effect of curriculum learning when applied to a large number of examples. We believe that when both the quantity and quality of examples are ensured, reasoning performance can be further improved, which will be a focus of our future work.

Potential Risks

Our work does not carry any obvious risks.

Acknowledgements

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11027–11052.

Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar. 2023. Chatgpt: Applications, opportunities, and threats. In *2023 Systems and Information Engineering Design Symposium (SIEDS)*, pages 274–279. IEEE.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. *arXiv preprint arXiv:2405.00200*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148.

Qi Guo, Leiyu Wang, Yidong Wang, Wei Ye, and Shikun Zhang. 2024. What makes a good order of examples in in-context learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14892–14904.

Guy Hacohen and Daphna Weinshall. 2019. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, pages 2535–2544. PMLR.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173.

Dan Iter, Reid Pryzant, Ruochen Xu, Shuohang Wang, Yang Liu, Yichong Xu, and Chenguang Zhu. 2023. In-context demonstration selection with cross entropy difference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1150–1162.

Jisu Kim and Juhwan Lee. 2024. Strategic data ordering: Enhancing large language model performance through curriculum learning. *arXiv preprint arXiv:2405.07490*.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. Mawps: A math word problem repository. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 1152–1157.

Bruce W Lee, Hyunsoo Cho, and Kang Min Yoo. 2023. Instruction tuning with human curriculum. *arXiv preprint arXiv:2310.09518*.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, and Wei Lu. 2024. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*.

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.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5011–5034.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing english math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences. *arXiv preprint arXiv:2302.11042*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. *arXiv preprint arXiv:2401.12087*.

Hadi Pouransari, Chun-Liang Li, Jen-Hao Rick Chang, Pavan Kumar Anasosalu Vasu, Cem Koc, Vaishaal Shankar, and Oncel Tuzel. 2024. Dataset decomposition: Faster llm training with variable sequence length curriculum. *arXiv preprint arXiv:2405.13226*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Fobo Shi, Peijun Qing, Dong Yang, Nan Wang, Youbo Lei, Haonan Lu, Xiaodong Lin, and Duantengchuan Li. 2024. Prompt space optimizing few-shot reasoning success with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1836–1862.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Yiming Tang and Bin Dong. 2024. Demonstration notebook: Finding the most suited in-context learning example from interactions. *arXiv preprint arXiv:2406.10878*.

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

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576.

Xin Wang, Yuwei Zhou, Hong Chen, and Wenwu Zhu. 2024. Curriculum learning: Theories, approaches, applications, tools, and future directions in the era of large language models. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1306–1310.

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

Noam Wies, Yoav Levine, and Amnon Shashua. 2024. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436.

Shan Xie, Man Luo, Chadly Daniel Stern, Mengnan Du, and Lu Cheng. 2024. Demoshapley: Valuation of demonstrations for in-context learning. *arXiv preprint arXiv:2410.07523*.

Xin Xu, Yue Liu, Panupong Pasupat, Mehran Kazemi, et al. 2024. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2023. Cumulative reasoning with large language models. *arXiv preprint arXiv:2308.04371*.

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

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Yaowei Zheng, Richong Zhang, Junhao Zhang, YeYanhan YeYanhan, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

A BREAK Dataset Description

BREAK is a dataset proposed by the Allen Institute(Wolfson et al., 2020). This work introduces the Question Decomposition Meaning Representation (QDMR), which breaks down a question into several sub-questions for solving and represents it as a sequence of steps. The dataset collects 83K question and QDMR pairs from several public datasets. To represent various questions as a unified sequence of steps, they customized 13 types of operations, converting the solution process for all questions into sequences of these operations. The specific operations and their examples are shown in Table 4. The decomposition and formalization of questions can be found in Figure 2.

Based on the BREAK dataset, we constructed an instruction set to analyze the problem-solving logic. Specific examples and explanations are provided in Table 5.

B Prompt Template

Table 6 shows the prompt templates used for fine-tuning problem-solving logic analysis.
Table 7–11 shows the full prompt example for in-context learning on the different benchmarks.

C Supplementary Details

Our experiments utilized the llama-factory (Zheng et al., 2024) project, which includes model fine-tuning and in-context learning. The CPU used in the experiments is an Intel(R) Xeon(R) Platinum 8358 CPU @ 2.60GHz, and the GPU is an NVIDIA Tesla A800 80G. The hyperparameters were set according to the default configuration file provided by llama-factory. The prompt length was set to 4096, and the maximum answer output length was set to 1024. To ensure output stability, the temperature was set to 0.01. In our study, we used ChatGPT to assist in writing code.

Operator	Question	Decomposition
Select	How many touchdowns were scored overall?	1. Return touchdowns 2. Return the number of #1
Filter	I would like a flight from Toronto to San Diego please.	1. Return flights 2. Return #1 from Toronto 3. Return #2 to San Diego
Project	Who is the head coach of the Los Angeles Lakers?	1. Return the Los Angeles Lakers 2. Return the head coach of #1
Aggregate	How many states border Colorado?	1. Return Colorado 2. Return border states of #1 3. Return the number of #2
Group	How many female students are there in each club?	1. Return clubs 2. Return female students of #1 3. Return the number of #2 for each #1
Superlative	What is the keyword, which has been contained by the most number of papers?	1. Return papers 2. Return keywords of #1 3. Return the number of #1 for each #2 4. Return #2 where #3 is highest
Comparative	Who are the authors who have more than 500 papers?	1. Return authors 2. Return papers of #1 3. Return the number of #2 for each of #1 4. Return #1 where #3 is more than 500
Union	Tell me who the president and vice-president are?	1. Return the president 2. Return the vice-president 3. Return #1, #2
Intersection	Show the parties that have representatives in both New York state and representatives in Pennsylvania state.	1. Return representatives 2. Return #1 in New York state 3. Return #1 in Pennsylvania state 4. Return parties in both #2 and #3
Discard	Find the professors who are not playing Canoeing.	1. Return professors 2. Return #1 playing Canoeing 3. Return #1 besides #2
Sort	Find all information about student addresses, and sort by monthly rental.	1. Return students 2. Return addresses of #1 3. Return monthly rental of #2 4. Return #2 sorted by #3
Boolean	Were Scott Derrickson and Ed Wood of the same nationality?	..... 3. Return the nationality of #1 4. Return the nationality of #2 5. Return if #3 is the same as #4
Arithmetic	How many more red objects are there than blue objects?	..... 3. Return the number of #1 4. Return the number of #2 5. Return the difference of #3 and #4

Table 4: Description of the 13 custom operations in the BREAK dataset.

Input
The input is a problem to be solved, such as: what flights are available tomorrow from denver to philadelphia?
Label
\\ *The label contains <operator> and <formal language>.* \\ *<operator> is an ordered set composed of the aforementioned custom operations.* \\ *<formal language> is the formalized language that provides a detailed description of each operator.* <operators>: ['select', 'filter', 'filter', 'filter'] <formal language>: ["SELECT['flights']", "FILTER['#1', 'from denver']", "FILTER['#2', 'to philadelphia']", "FILTER['#3', 'if available']"]

Table 5: Examples and Explanation of Instruction Sets Based on the BREAK Dataset

Prompt
You are a helpful assistant. Please break down in order the operations <operations> required to solve the following problems, and the process of solving the problem according to the operations <programs>: what flights are available tomorrow from denver to philadelphia?
Label
<operators>: ['select', 'filter', 'filter', 'filter'] <formal language>: ["SELECT['flights']", "FILTER['#1', 'from denver']", "FILTER['#2', 'to philadelphia']", "FILTER['#3', 'if available']"]

Table 6: Fine-tuning prompts for problem-solving logic analysis

Prompt
System prompt Please provide the answer in the following format: "The final answer is <answer>" **User input** question: Being his favorite, he saved checking on the grapevines for his last stop. He was told by 235 of the pickers that they fill 100 drums of raspberries per day and 221 drums of grapes per day. How many drums of grapes would be filled in 77 days? answer: Equation is (221.0 * 77.0). The final answer is 17017.0 question: Tiffany was collecting cans for recycling. On Monday she had 4 bags of cans. The next day she found some more bags worth of cans. If she had a total of 6 bags altogether, how many bags did she find on the next day? answer: Equation is (6.0 - 4.0). The final answer is 2.0 question: After a typhoon, 13 trees in Haley's backyard died. If she had grown 3 trees initially, how many more trees died in the typhoon than those that survived? answer: Equation is (13.0 - (3.0 - 13.0)). The final answer is 23.0 question: Brenda's mother made cookies for 5 people. She prepared 22 cookies but had to throw away 17 cookies. If each of them had the same number of cookies, how many did each of them have? answer: Equation is ((22.0 - 17.0) / 5.0). The final answer is 1.0 question: Haley grew 9 trees in her backyard. After a typhoon 4 died. Then she grew 5 more trees. How many trees does she have left?
Number of problem-solving logic step: 4->5->6->7->7

Table 7: Full prompt example for in-context learning on the SVAMP dataset

Prompt
System prompt
Please provide the answer in the following format: "The final answer is <answer>"
User input
question: What is the probability of rolling one six-sided dice, and getting a different number on each die? A)1/12 B)1/3 C)1 D)5/9 E)7/18 answer: Let's think step by step. MAGOOSHOFFICIAL SOLUTION. For the first die, we can roll any one of six numbers. For the second die, we can roll any number save for the number we rolled on the first die, giving us 5 possibilities. For the third die, we can roll four different numbers (we can't roll the number we rolled on the first or second die). 6 possibilities out of 216 total possibilities. 6/6 = 1. The final answer is C. question: If q is the square of a positive integer, which of the following must be equal to the square of the next positive integer? A)$\sqrt{n+1}$ B)$n+1$ C)n^2+1 D)$q+2\sqrt{q}+1$ E)n^2+2n+1 answer: Let's think step by step. If q is the square of a positive integer, then $q=(x)^2$ where x is a positive integer. To calculate, $(x+1)^2=x^2+2x+1$, which is $q+2\sqrt{q}+1$. This should be D. The final answer is D. question: If Tim had lunch at $50 and he gave 20% tip, how much did he spend? A)$60.00 B)$35.42 C)$60.60 D)$21.56 E)$78.45 answer: Let's think step by step. The tip is 20% of what he paid for lunch. Tip = 20% of 50.00 = $10.00. Total spent = 50.00 + 10.00 = $60.00. The final answer is A. question: Carl is facing very difficult financial times and can only pay the interest on a $10,000 loan he has taken. The bank charges him a quarterly compound rate of 4%. What is the approximate interest he pays annually? A)$1600 B)$2000 C)$2150 D)$2500 E)$12000 answer: Let's think step by step. The bank charges a 4% quarterly compounded annual rate. Per quarter rate is (16/4)% = 4%. Thus, the quarterly compounded interest will be slightly more than $1600. The final answer is A. question: A shopkeeper employed a servant at a monthly salary of 1500. In addition to it, he agreed to pay him a commission of 15% on the monthly sale. How much sale in Rupees should the servant do if he wants his monthly income as 6000? A)30000 B)415000 C)31500 D)50000 E)None of these
Number of problem-solving logic step:
2->3->4->5->6

Table 8: Full prompt example for in-context learning on the AQuA dataset

Prompt
System prompt Please provide the answer in the following format: "The final answer is <answer>" **User input** question: A shopkeeper bought 150 packets of milk. Each packet contained 250 ml of milk. If one fluid ounce is equal to 30 ml, how many ounces of milk did he buy? nanswer: Let's think step by step. If the shopkeeper bought 150 packets of milk, each packet containing 250ml of milk, all the packets had a total of 250*150 =«150*250=37500»37500ml.Since one ounce equal 30 ml, the total amount of milk that the shopkeeper bought in oz is 37500/30=«37500/30=1250»1250 oz of milk. The final answer is 1250 question: Twenty gallons of tea were poured into 80 containers. Geraldo drank 3.5 containers. How many pints of tea did Geraldo drink? answer: Let's think step by step. 20 gallons = 160 pints. 160/80 = «160/80=2»2 pints.3.5 * 2 pints = «3.5*2=7»7 pints. Geraldo drank 7 pints of tea. The final answer is 7 question: During the holidays, Lance works as a merchandiser. He works 35 hours a week, spread equally over 5 workdays. If Lance earns $9 an hour, how much does he make on each workday? answer: Let's think step by step. Lance works 35 / 5 = «35/5=7»7 hours a day. So he makes $9 x 7 = $«9*7=63»63 on each workday. The final answer is 63 question: A snack machine accepts only quarters. Candy bars cost ¢25, each piece of chocolate costs ¢75, and a pack of juice costs ¢50. How many quarters are needed to buy three candy bars, two pieces of chocolate, and one pack of juice? answer: Let's think step by step. Three candy bars cost ¢25 x 3 = ¢«25*3=75»75. Two pieces of chocolate cost ¢75 x 2 = ¢«75*2=150»150. So, the total amount needed to buy those is ¢75 + ¢150 + ¢50 = ¢«75+150+50=275»275. Since a quarter is equal to ¢25, therefore ¢275/¢25 = «275/25=11»11 quarters are needed. The final answer is 11 question: Mark makes custom dog beds. A bed for a Rottweiler takes 8 pounds of stuffing, a bed for a chihuahua takes 2 pounds of stuffing, and a bed for a collie takes the average amount of stuffing between the first two kinds of beds. How many pounds of stuffing does Mark need to make 4 chihuahua beds and 3 collie beds?
Number of problem-solving logic step: 5->6->7->8->8

Table 9: Full prompt example for in-context learning on the Gsm8k dataset

Prompt
System prompt Please provide the answer in the following format: "The final answer is <answer>" **User input** question: What is the only was to recover from exhaustion? A. mediate B. have rest C. stay in bed D. run out of steam E. go to sleep answer: B question: Google Maps and other highway and street GPS services have replaced what? A. united states B. mexico C. countryside D. atlas E. oceans answer: D question: You can share files with someone if you have a connection to a what? A. freeway B. radio C. wires D. computer network E. electrical circuit answer: D question: If a person isn't able to pay their bills what must they do? A. know everything B. acknowledgment C. make more money D. throw a party E. spare time
Number of problem-solving logic step: 1->2->3->3

Table 10: Full prompt example for in-context learning on the ComSenQA dataset

Prompt
System prompt Please provide the answer in the following format: "The final answer is yes or no" **User input** question: Can you buy Casio products at Petco? answer: Casio is a manufacturer of consumer electronics and watches. Petco is a chain store that sells pet supplies like food, bowls, litter, toys, cages and grooming equipment. The final answer is no question: Did Clark Gable appear in any movies scored by John Williams? answer: Clark Gable died in 1960. John Williams scored his first movie in 1961. The final answer is no question: Could a dandelion suffer from hepatitis? answer: Only creatures that contain a liver can suffer from hepatitis. The liver is an organ only found in vertebrates. Vertebrates exist in the kingdom Animalia. Dandelions are plants in the kingdom Plantae. The final answer is no question: Did Mozart ever buy anything from Dolce & Gabbana?
Number of problem-solving logic step: 2->3->4->4

Table 11: Full prompt example for in-context learning on the StrategyQA dataset