On the Role of Chain of Thoughts in the Long In-Context Learning

Anonymous ACL submission

Abstract

In-Context Learning (ICL) has emerged as a powerful paradigm for adapting Large Language Models (LLMs) to new tasks without gradient updates. While advances in long-context models have enabled a shift from few-shot to many-shot ICL, achieving performance comparable to fine-tuning, research has largely focused on classification tasks. Therefore we target to address the underexplored behavior of Chain-of-Thought (CoT) prompting in many-shot scenarios, explaining the shift from studies in how to select appropriate ICL examples to how to enable LLMs to evolve at test-time. We present a comprehensive analysis of many-shot in-context CoT learning, uncovering behavioral differences between reasoning-oriented and non-reasoning oriented LLMs. Our findings show that with both types of models, there is a fundamental difference from earlier studies in the ICL setting. Demonstration selection based on input similarity, which is a common heuristic in ICL, becomes ineffective under the CoT paradigm. Counterintuitively, our experiments show that the quality of the reasoning chain, as measured by its ground-truth correctness, is not the primary factor for success. Instead, we observe a consistent performance gap where model-self-generated CoTs with incorrectness outperform those with human-verified, correct reasoning. This suggests that the success of many-shot CoT prompting is driven more by alignment with the model's internal reasoning patterns than by the objective quality of the demonstrations. These observations indicate that LLMs are more effective at learning from their own "experiences" than from ground-truth reasoning or externally constructed examples, which they may find harder to interpret. This insight also explains why similarity-based selection strategies lose their effectiveness in CoT settings. Additionally, our experiments reveal that, unlike previous many-shot ICL studies, the order of CoT demonstrations significantly affects performance. We observe a fluctuating or increasing variance along the number of shots when the demonstration order is shuffled. This points to the critical role of ordering in the effectiveness of LLM test-time learning.

1 Introduction

In-Context Learning (ICL), where Large Language Models (LLMs) are prompted with a sequence of input-output demonstrations and asked to produce predictions for new inputs without any gradient updates, has gained significant attention. Research has extensively investigated its benefits (Sorensen et al., 2022; An et al., 2023; Mavromatis et al., 2023) and underlying mechanisms (Min et al., 2022; von Oswald et al., 2023; Deutch et al., 2024). A substantial body of work has focused on enhancing ICL, including further pre-training models for improved ICL capability, developing strategies for demonstration selection, and exploring the transition from few-shot to many-shot learning. While early work focused on few-shot ICL, recent advances in scaling context windows have made it possible to explore many-shot ICL, where dozens to hundreds or even more demonstrations can be provided, allowing performance comparable to fine-tuning (Agarwal et al., 2024; Bertsch et al., 2025; Baek et al., 2025). However, the majority of these studies focus on classification-oriented ICL, where tasks are relatively shallow and answers are directly inferred from patterns in the demonstrations.

In parallel, there has been growing interest in chain-of-thought (CoT) prompting, a technique that improves reasoning tasks by encouraging models to generate intermediate reasoning steps before arriving at a final answer (Kojima et al., 2022). While CoT has shown strong performance in few-shot settings (Zhang et al., 2023; Wei et al., 2022; Luo et al., 2023), its behavior in many-shot in-context learning remains underexplored. Crucially, in the case of ICL with CoT, most prior work directly applies the few-shot paradigm without delving deeply into the underlying mechanisms. This raises several critical questions.

1. What happens in the many-shot scenario: does performance scale monotonically, or does it plateau or even degrade?

2. Is in-context CoT fundamentally different from ICL with a single label?

3. What correlating factors govern the effectiveness of many-shot CoT, and can previous demonstration selection strategies (e.g., based on semantic similarity) be applied directly?

This gap is significant in light of the growing context lengths supported by LLMs and the emerging concept of test-time scaling and DeepResearch. Scaling in test-time enables LLMs to refine their responses without parameter updates during inference, through paradigms such as sequential revision and parallel sampling (Snell et al., 2025). Prior works have explored inference-time alignment methods with ICL (Lin et al., 2024; Li et al., 2025). Li et al. (2025) proposes parallel sampling for sequential refining and shows that increasing the search width of sampled responses in in-context consistently enhances the performance, demonstrating the potential of incorporating in-context demonstrations in enhancing the LLM capability during inference. Yet it remains unclear whether in-context CoT similarly benefits from longer contexts, or whether it introduces new challenges due to the complexity of reasoning chains.

In this work, we present a comprehensive analysis of many-shot in-context CoT learning, comparing its behavior with traditional classification-based ICL and evaluating its effectiveness under extended context lengths. Our study identifies several fundamental differences between CoT and classification ICL under many-shot settings. We find that, in contrast to many-shot ICL in tasks without involving CoT (Bertsch et al., 2025; Baek et al., 2025), the performance of in-context CoT is highly sensitive to demonstration ordering and selection. For demonstration selection, unlike traditional ICL, similarity no longer serves as a reliable signal, highlighting the need for further investigation specifically for in-context CoT. These findings suggest that many-shot CoT learning is governed by different dynamics than those observed in traditional ICL. To address these challenges, we investigate factors correlated with in-context CoT performance and find that constructing LLM-aligned CoT demonstrations stabilizes its performance. Our results show that the quality of the provided CoT is unexpectedly not a critical factor. We observe a clear hierarchy where demonstrations with self-generated, incorrect CoTs lead to the best performance outperforming ground-truth human-verified CoTs. Adopting CoTs generated by a stronger, more advanced LLM, on the other hand, results in a lower accuracy. This suggests that the mechanism behind many-shot CoT is not merely about providing higher-quality examples, but is related to the alignment to LLM.

2 Mysteries of many-shot ICL

2.1 Experiment Setup

Tasks. Previous studies in many-shot (Li et al., 2024; Bertsch et al., 2025) lacks exploration in the reasoning tasks. Experiments are conducted with a diverse type of tasks, including traditional classification tasks (i.e., SuperGLUE (Wang et al., 2019) with a narrow label space; NLU (nlu, 2021), TREC (Hovy et al., 2001) and BANKING77 (Casanueva et al., 2020) with significantly larger label space), mathematical reasoning (i.e., GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021)).

ICL Settings. For the reasoning tasks, the inclusion of CoT is natural. Hoping to provide the complete picture also for the classification tasks, we also adopt the LLM-generated CoT for the classification tasks.

- **Traditional ICL.** An instance consists of an input-output pair (x,y). With n in-context demonstrations provided, the LLM processes the input as $\texttt{LLM}(x' \mid \{(x_i, y_i)\}_{i=1}^{n})$ to generate y'. Since the token length of individual instances is relatively small, current open-source LLMs can easily handle hundreds or even thousands of examples during evaluation.

- **In-context CoT.** An instance consists of an input-CoT-output triplet (x, C, y). With n in-context demonstrations provided, the LLM processes the input as $\texttt{LLM}(x' \mid \{(x_i, C_i, y_i)\}_{i=1}^{n})$ to generate y'. Since the token length of individual instances can be significantly larger, depending on the token length of C, current open-source LLMs are typically limited to evaluating only around hundreds of examples.

LLMs Studied. Building on previous studies that highlight the limitations of long-context LLMs in ICL (Li et al., 2024; Bertsch et al., 2025) and given the recent advancements in instruction-tuned models with extended context windows in about 130K tokens, we conduct experiments with LLaMA 3.1 (`Llama-3.1-8B-Instruct`), LLaMA 3.3 (`Llama-3.3-70B-Instruct`) (MetaAI, 2024), Qwen 2.5 (7B) (`Qwen2.5-7B-Instruct`), Qwen 2.5 (14B) (`Qwen2.5-14B-Instruct`), Qwen 3 (8B) (`Qwen3-8B`), Qwen 3 (14B) (`Qwen3-14B`) (Qwen et al., 2025), enabling analysis across different LLM architectures and model sizes. To enable the processing of long context to the 131k token level for the Qwen family, we modified the config file and add the `rope_scaling` fields.

Unlike prior studies that focus primarily on classification tasks with constrained decoding (Bertsch et al., 2025), we adopt a generative framework for both classification and generation tasks. Specifically, we formulate all tasks as text generation problems and evaluate model outputs using exact match against reference answers or labels. This approach aligns with recent trends in LLM research, where the emphasis has increasingly shifted toward open-ended generation settings.

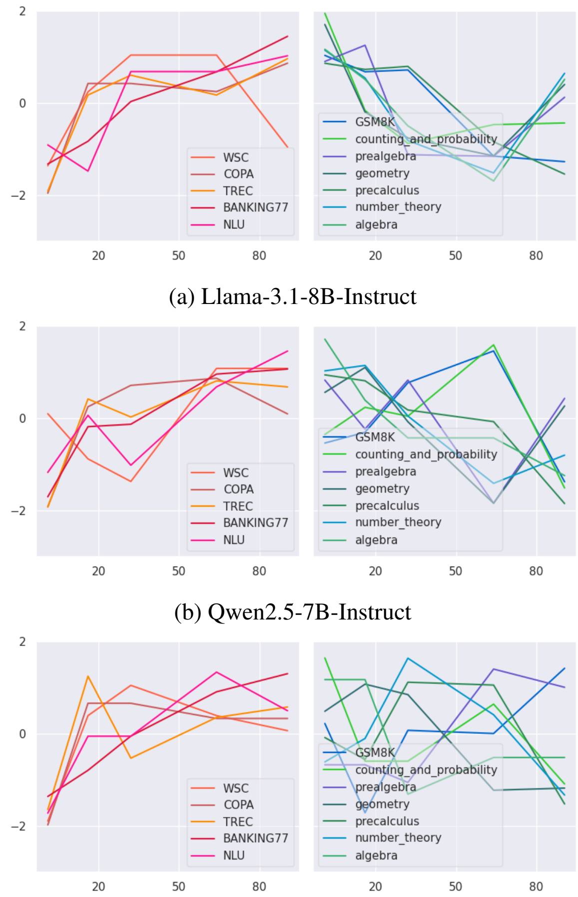
In many-shot in-context CoT learning, the number of tokens per demonstration can be substantially larger than in traditional many-shot ICL due to the length of CoT reasoning. For instance, when comparing geometry task to BANKING77, the average demonstration length in the former is 30 times longer, averaged across the training set. To maintain consistency with in-context CoT scaling, we prompt with approximately 100-shot in-context demonstrations in the following studies.

In addition, empirical results in the following subsection show that even with the Qwen3 family, model performance declines sharply beyond a certain number of tokens. This suggests limited benefit in further increasing the number of in-context demonstrations under current model constraints. Under these considerations, our analysis in performed on the scope of about a hundred demonstrations.

2.2 Results

To enable a direct comparison of performance across tasks with varying accuracy ranges, we normalize all results, as illustrated in Figure 1. Detailed results are provided in Appendix B.

As shown in the figure, there is a significant difference between the two types of tasks: classification tasks exhibit a consistent pattern of steady improvement as the number of demonstrations increases, whereas math reasoning tasks show fluctuating or even declining performance. As pointed



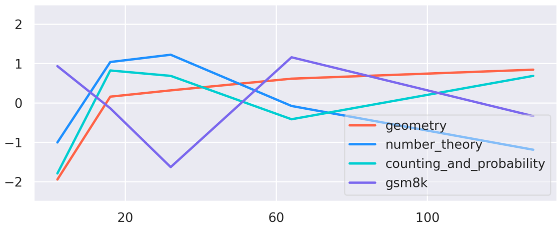
(a) Llama-3.1-8B-Instruct

(b) Qwen2.5-7B-Instruct

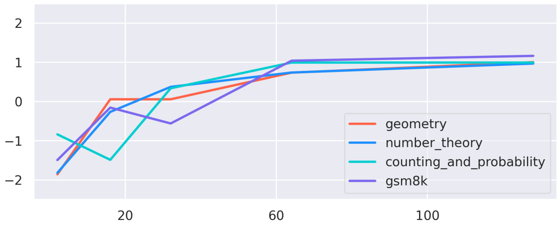
(c) Qwen2.5-14B-Instruct

Figure 1: Performance comparison using normalized results between classification tasks (in warm colors) and math reasoning tasks (in cool colors). The x-axis represents normalized accuracy, while the y-axis indicates the number of in-context demonstrations.

Figure 2: Performance on randomly sampled subset of math reasoning tasks using Llama-3.3-70B-Instruct (left) and QwQ-32B (right).



(a) Qwen3-8B



(b) Qwen3-14B

Figure 3: Performance on randomly sampled subset of math reasoning tasks using Qwen 3 family.

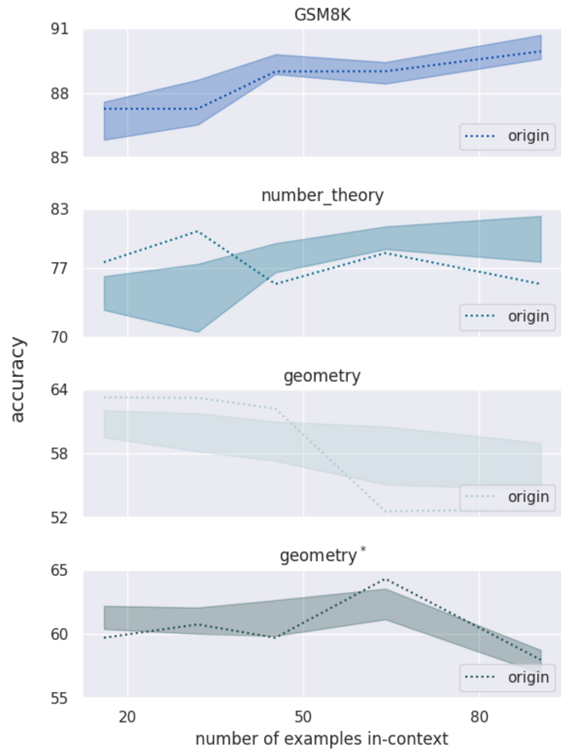


Figure 4: The original performance with the error band computed across five random orders shuffled with a unified set of random seeds under Qwen 2.5 (14B). `geometry`* is evaluated with another randomly sampled in-context CoT demonstration set.

out by Li et al. (2024), LLMs often struggle to learn effectively from long in-context examples.

With the advancement of LLMs, subsequent studies have shown that learning from long-context classification tasks has become increasingly effective (Agarwal et al., 2024; Bertsch et al., 2025). While LLMs demonstrate strong many-shot learning capabilities in classification tasks, their performance in reasoning-intensive tasks using in-context chain-of-thought (CoT) remains limited. This limitation persists even in recent LLMs: both the Qwen 2.5 and LLaMA 3.1 families show difficulty in handling long-context CoT settings. To further investigate, we evaluate a larger model, LLaMA-3.3-70B. As shown in Figure 2, even with this increased parameter size, the model is still unable to effectively learn from the provided demonstrations.

The recently released Qwen 3 family demonstrates a general trend of steadily increasing performance, as shown in Figure 3. To explore the limits of their capability, we leveraged its extended context window, pushing to the maximum token limit of 131k. This allowed for the inclusion of up to 256 in-context demonstrations in the geometry task, with each demonstration comprising approximately 450 tokens. However, we observe a severe and progressive performance degradation beyond a certain threshold of demonstrations. For instance, on the geometry task, the accuracy of Qwen3-14B progressively degrades from 64.92% (with 128 demonstrations) to 55.74% (with 181 demonstrations), before collapsing to 12.32% (with 256 demonstrations). A similar pattern is observed with LLaMA 3.1 and Qwen 2.5 (8B), while Qwen 2.5 (14B) exhibits a less severe, though still present, fluctuation in performance.

Consequently, despite the Qwen 3 family's robust overall performance, our experiments show that most LLMs struggle to effectively learn from a large number of in-context Chain-of-Thought demonstrations. This consistent failure mode across model families raises a research question: can any strategies be adopted to enhance CoT performance in long-context reasoning tasks?

2.3 Is LLM truly incapable to be benefitted from in-context CoT demonstrations?

While the recently released models no longer struggle with the large space classification tasks assessed by Li et al. (2024), new challenges arrived with the inclusion of the reasoning chain. To further investigate the effectiveness of in-context CoT, we conduct a focused analysis on three tasks (i.e., GSM8K, geometry and number_theory) with Qwen 2.5 (14B). Our analysis proceeds along two strategies:

1. **Demonstration Order Sensitivity:** We randomly shuffle the order of demonstrations five times to examine whether an ordering that facilitates better in-context CoT exists.

2. **Demonstration Choice Sensitivity:** We randomly sample alternative sets of in-context examples to assess whether performance improvements can be attributed to the choice of demonstrations.

The results are shown in Figure 4, presenting the original performance trends and the error band of mean$\pm$standard deviation across five accuracies with shuffled orders. To ensure experimental fairness, a fixed set of seeds is randomly sampled to perform the order shuffles. The same seeds are consistently applied across all configurations, spanning different LLMs, numbers of in-context examples, and tasks to maintain controlled variability.

For the GSM8K and number_theory tasks, an increasing trend is observed with the first strategy. Notably, although the original trend for number_theory was decreasing, both the upper-bound performance and the average performance across five orderings show an increasing trend. These indicate a sensitivity to the order of demonstrations.

In the geometry task, the same trend is observed after the initial attempt of applying the second strategy, continuing up to 64 demonstrations. The subsequent decline in performance can be attributed to the context length exceeding 40k tokens, which necessitates the use of RoPE scaling. This indicates a sensitivity to the selection of demonstrations.

3 Do LLMs reason with in-context CoT?

A key question in analyzing the performance of in-context CoT reasoning will be whether its success directly correlates with the model's task understanding or whether other factors play a more significant role. To investigate this, we create two 4-option multiple-choice question answering (MCQA) tasks and found that in-context CoT may correlate more with data distribution than task understanding in Appendix A.

3.1 Correlating Factor Influencing In-Context CoT Success

Tasks. We follow the previous settings and select a subset of math reasoning tasks for further analysis, including geometry, number theory, and GSM8K. With wider application of LLMs into real-life tasks, especially with the development of Deep Research (Huang et al., 2025; Yu et al., 2025), the rationale behind comprehensive reasoning tasks also matters. Therefore, apart from the math problems and classification tasks, we additionally studied LLM performance that required narrative reasoning with DetectiveQA (Xu et al., 2025). Since DetectiveQA provides the corresponding evidence and the reasoning chain for deriving the answer based on the evidence. For each instance, the evidence is provided as part of the question. The corresponding CoT will be the derivation labelled with "-1".

LLMs and Tasks Studied We evaluate with LLaMA 3.1 and Qwen 2.5 (14B) on three math reasoning benchmarks (i.e., GSM8K, number_theory, and geometry) and evaluate with Qwen 3 (8B) and Qwen 3 (14B) on number_theory and DetectiveQA.

Model-CoT alignment. We investigate whether the efficacy of in-context Chain-of-Thought (CoT) learning is more significantly influenced by the logical quality of the reasoning or its alignment with the LLM's own generative distribution. To this end, we generate CoT demonstrations by prompting the LLM on the training set, rather than using the dataset's ground-truth CoT. These generated CoT are then used as the in-context examples during evaluation over the test set.

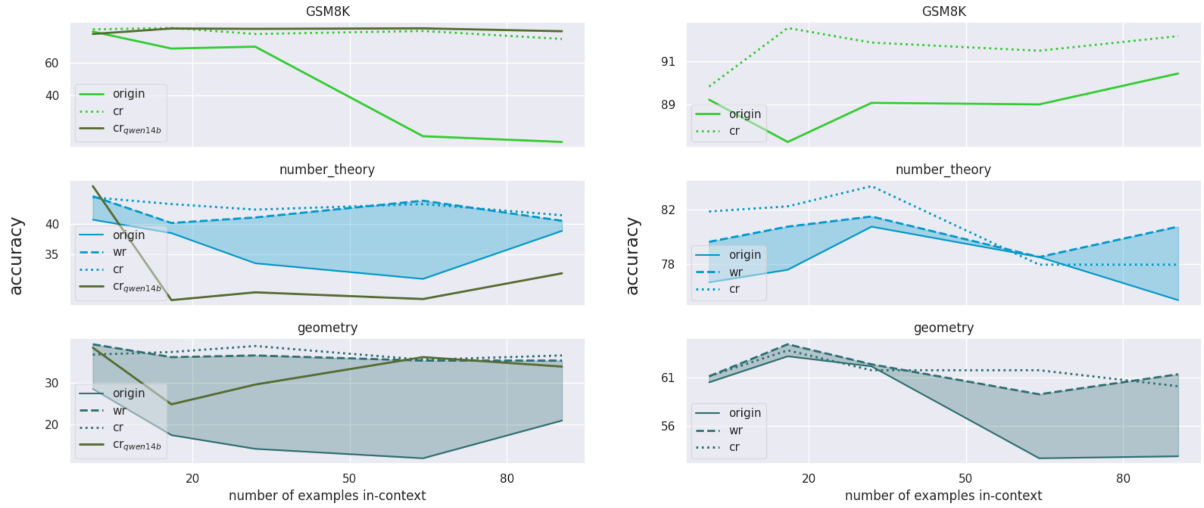
To investigate whether CoT quality or distributional alignment plays a more significant role in performance, we construct three distinct demonstration sets to isolate these factors:

1. The Correct Set (cr): Samples where the model's generated answer is correct.

2. The Incorrect Set (wr): Samples where the model's generated answer is incorrect.

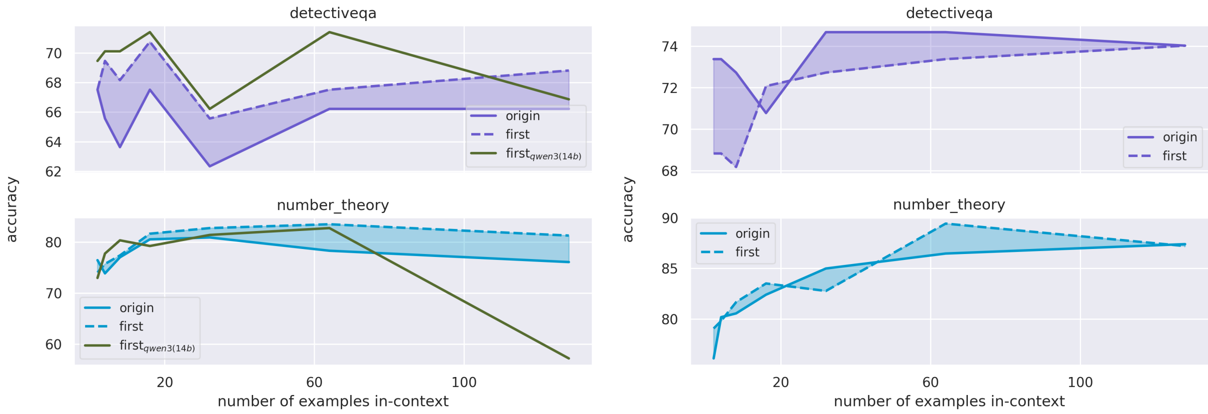
3. The First Set (first): The initial generation for each instance, regardless of accuracy.

Each LLM is prompted 10 times per training instance using a temperature of 1.0 for diversity. During the 10 times of prompting, if the predicted answer is correct, we include the corresponding CoT and generated answer in the cr set; otherwise, it is placed in the wr set. The first generation is included in the first set. These resulting demonstration sets are compared against the original CoT (i.e., origin) provided within the datasets.



(a) Llama-3.1-8B-Instruct (b) Qwen2.5-14B-Instruct

Figure 5: Performance of two sets of self-generated in-context CoT, including the set filtered with only correct answer(cr) and the set filtered with only wrong answer(wr). cr$_{\text{qwen14b}}$ is prompting the LLaMA model with the in-context CoT generated by Qwen 2.5 (14B).



(a) Qwen3-8B (b) Qwen3-14B

Figure 6: Performance of the first set of self-generated in-context CoT. cr$_{\text{qwen3(14b)}}$ is prompting the Qwen 3 (8B) model with the in-context CoT generated by Qwen 3 (14B).

Due to the high accuracy of both LLMs on GSM8K, it is difficult to obtain incorrect outputs even at a high temperature. Thus, the wr set is only constructed for number_theory and geometry.

Result Surprisingly, as illustrated in Figure 5, the wr set with wrong answers and presumably flawed reasoning always outperforms the original CoT and performs comparably to the cr set across both LLMs and both tasks. This shows the effectiveness of having LLM-aligned in-context CoT. Additionally, with the self-generated CoT, both LLMs suffer significantly less from the sudden drop and great fluctuation issues, especially for LLaMA 3.1.

Moreover, the use of self-generated CoT significantly mitigates the issues of performance instability and sudden accuracy drops observed with the origin, an effect particularly pronounced for LLaMA 3.1. Collectively, these results suggest that the distributional characteristics of the in-context examples (i.e., their alignment with the LLM's own generative patterns) exert a more substantial influence on stable CoT prompting than the conventional metric of quality, defined here as the presence of a correct final answer.

In the meantime, since the wr and cr sets were constructed by sampling model outputs, an instance can have all correct or all incorrect answers across all runs, risking an empty instance. To ensure robustness, we perform the analysis using the first set with guaranteed no emptyness.

Results in figure 6 reinforce the initial finding. The first set again outperforms the original CoT in both the number theory and DetectiveQA tasks, suggesting data distribution is a more influential factor than quality for stable in-context CoT prompting.

Does "better" CoT give better and stable performance? To further assess the role of CoT quality, we investigate whether better CoT from a better-performing model can improve the performance of a weaker one. Specifically, we prompt LLaMA 3.1 using CoT generated by Qwen 2.5 (14B) in Figure 5 and Qwen 3 (8B) using CoT generated by Qwen 3 (14B) in Figure 6, where Qwen 2.5 (14B) and Qwen 3 (14B) both show a better or comparable performance with the baseline.

As shown in the olive line in Figure 5 and 6, while LLaMA 3.1 and Qwen 3 (8B) does benefit from higher-quality CoT with a performance increase in geometry and DetectiveQA at the beginning and at certain shots of in-context examples, the model still suffers significantly from instability, including occasional sudden performance drops and great fluctuations. This implies that while higher-quality CoT may enhance performance, but likely to result in greater instability without good in-context data distribution (i.e., not well-aligned with the evaluation LLM). This further reinforces our finding that data distribution is a more crucial factor in successful in-context CoT prompting.

3.2 Rethinking the role of similarity

Previous research in in-context learning has shown that retrieving semantically similar examples often enhances model performance (Liu et al., 2022; Wu et al., 2023; Kapuriya et al., 2025). To further investigate this claim, we include the BANKING77 classification task as a control experiment. Specifically, we construct two unified sets of in-context examples: one comprising the most semantically similar examples and the other comprising the most dissimilar examples to the test set. Similarity is measured by computing the cosine similarity between question embeddings, averaged across the entire test set with a sentence transformer (`all-mpnet-base-v2`). Approximately 250 samples are retrieved from the training set to form the candidate pool for constructing the similar and dissimilar example sets. We evaluate performance on both the BANKING77 task and the reasoning tasks introduced in Section 2.3 with Qwen 2.5 (14B).

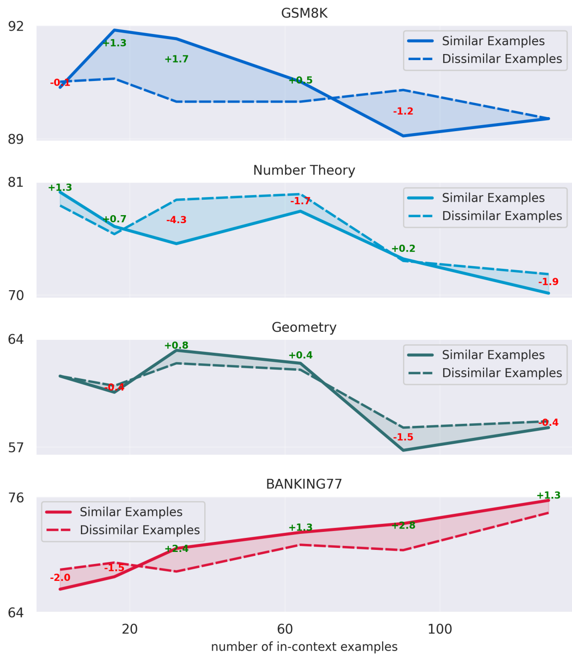


Figure 7: Performance with similarity(sim) and dissimilar(dis) sets with Qwen 2.5 (14B). The area between the two sets is filled with colors, indicating the relative performance at each point.

The results are presented in Figure 7. Since our retrieval strategy is based on the global similarity to the full test set rather than per-instance similarity, the benefits of similarity-based retrieval may not be as apparent in few-shot ICL. However, in settings with more than 20 in-context examples, the similar set consistently yields a better performance over the dissimilar in BANKING77, with the area in between highlighted in green. This aligns with prior findings in classification-based ICL. In contrast, we observe the opposite trend for the three reasoning tasks. With the increasing number of in-context CoT provided, the dissimilar set consistently outperforms the similar set, showing prior findings in ICL cannot be extended to in-context CoT.

4 Direction Towards Test-time Scaling.

With demonstrated effectiveness with self-generated CoT, learning is demonstrated within the provided instance. This creates the question of how to enable better learning from the provided demonstrations. Bertsch et al. (2025); Baek et al. (2025) found that the impact of demonstration ordering diminishes as the number of demonstrations increases. Building on this insight, we further explore order sensitivity in in-context chain-of-thought (CoT) prompting. Specifically,

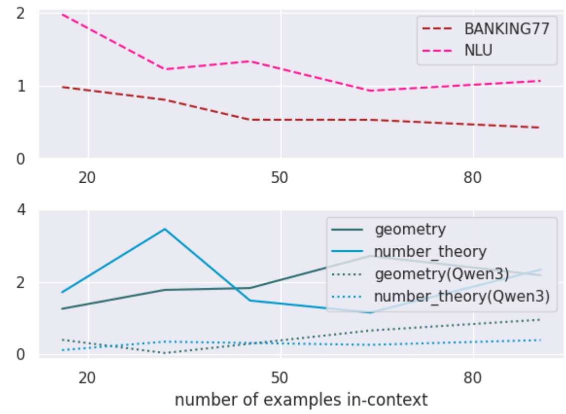


Figure 8: Standard deviation across five order sampling on Qwen 2.5 (14B) and Qwen 3 (14B).

we calculate the standard deviation across five accuracy scores obtained from randomly shuffled demonstration orders. Consistent with prior findings, classification tasks (e.g., NLU and BANKING77) show a clear pattern, the standard deviation in performance decreases as more demonstrations are added. Aligning to previous understanding, classification tasks can be resolved by locating similar instances, leading to the stability of performance after providing a sufficiently large ICL examples, no matter how the order of demonstrations is.

In contrast, for reasoning tasks, the standard deviation either fluctuates unpredictably or gradually increases with more demonstrations. This implies that, with complex problems necessitating reasoning, order is one of the important factors affecting how LLM can learn during test-time. This also align to human learning, where the chapters of the textbook are arranged in a specifically designed order to foster better learning and understanding.

5 Related Works

5.1 Many-shot ICL

With studies enabling LLMs to handle longer context lengths (Peng et al., 2024; Han et al., 2024; Ding et al., 2024), Agarwal et al. (2024) introduce the concept of many-shot ICL, which incorporates a significantly larger number of in-context demonstrations. Their results show a comparable performance to fine-tuning across various types of tasks. Subsequent studies (Li et al., 2024; Bertsch et al., 2025) investigate the effectiveness of many-shot ICL in open-source LLMs and examine its distinct characteristics compared to few-shot ICL. Bertsch et al. (2025); Baek et al. (2025) report reduced sensitivity to demonstration selection, and Baek et al. (2025) highlight increased vulnerability to noisy examples in complex tasks. However, these works primarily focus on traditional classification-based ICL without exploring in-context CoT reasoning. Given the growing attention to the reasoning capabilities of LLMs and the demonstrated effectiveness with a large number of in-context demonstrations at test-time in enhancing model performance (Li et al., 2025), it becomes crucial to understand how many-shot CoT behaves under long-context settings. Our study provides a comprehensive evaluation of many-shot in-context CoT and investigates their unique characteristics, revealing deviations from the previously observed patterns in traditional many-shot ICL.

5.2 Chain-of-Thought

Prior studies have focused on modifying and enhancing the Chain-of-Thought (CoT) prompting paradigm to improve reasoning performance in large language models. Program of Thoughts (PoT) (Chen et al., 2023) introduces structured programming to represent the reasoning process more systematically. Tree-of-Thoughts (ToT) (Yao et al., 2023) proposes a tree-structured reasoning framework that enables the model to explore different reasoning paths. rStar-Math (Guan et al., 2025) decomposes complex reasoning problems and explores diverse reasoning trajectories using Monte Carlo Tree Search (MCTS) at test time, achieving significant improvements on mathematical reasoning benchmarks. In the meantime, only a few studies have explored the application of CoT in ICL. Dr.ICL (Luo et al., 2023) extends retrieval-augmented ICL to CoT prompting, demonstrating notable gains in mathematical reasoning tasks. However, these works typically operate under one-shot or few-shot ICL settings, leaving the potential of in-context CoT across extended context lengths largely underexplored. Our work addresses this gap by investigating how CoT prompting scales with increasing context length.

6 Conclusion

In this study, we provided a thorough investigation into the behavior of in-context CoT learning in extended context settings and its comparison with traditional ICL approaches for classification tasks. Our analysis uncovered several unique challenges faced by in-context CoT learning, particularly its sensitivity to factors such as demonstration ordering and selection. This contrasts with prior findings in many-shot ICL, where such sensitivities are found to be less pronounced. Notably, we find that retrieving similar demonstrations does not enhance in-context CoT performance, diverging from established results in classification-based ICL.

Our empirical findings also demonstrate that task performance in many-shot in-context CoT is more influenced by the underlying data distribution more than task comprehension. Further experiments show that aligning in-context CoT demonstrations with LLMs' internal priors and learned reasoning trajectory can lead to more stabilized and consistent performance. Our work highlights the difference between in-context CoT with previous studies and the need for tailored strategies in leveraging in-context CoT learning, helping to lay the ground for further exploration of its potential and limitations.

Limitations

Due to the computational cost and performance limitations of LLMs in long in-context CoT reasoning, our study is limited to approximately 100 examples. While LLMs like Qwen 2.5 and LLaMA 3.1 can handle up to 131K and 128K context tokens, respectively, their performance in in-context CoT reasoning declines gradually beyond a certain threshold of context tokens, making exploring beyond 100 shots in this setting insignificant.

References

2021. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, Lecture Notes in Electrical Engineering, pages 165–183. Springer.

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie C.Y. Chan, Biao Zhang, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. In *ICML 2024 Workshop on In-Context Learning*.

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, Toronto, Canada. Association for Computational Linguistics.

Jinheon Baek, Sun Jae Lee, Prakhar Gupta, Geunseob Oh, Siddharth Dalmia, and Prateek Kolhar. 2025. Revisiting in-context learning with long context language models. *Preprint*, arXiv:2412.16926.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2025. In-context learning with long-context models: An in-depth exploration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12119–12149, Albuquerque, New Mexico. Association for Computational Linguistics.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

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

Gilad Deutch, Nadav Magar, Tomer Natan, and Guy Dar. 2024. In-context learning and gradient descent revisited. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1017–1028, Mexico City, Mexico. Association for Computational Linguistics.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. LongroPE: Extending LLM context window beyond 2 million tokens. In *Forty-first International Conference on Machine Learning*.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *Preprint*, arXiv:2501.04519.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2024. LM-infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3991–4008, Mexico City, Mexico. Association for Computational Linguistics.

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. *NeurIPS*.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, Kun Shao, and Jun Wang. 2025. Deep research agents: A systematic examination and roadmap. *Preprint*, arXiv:2506.18096.

Janak Kapuriya, Manit Kaushik, Debasis Ganguly, and Sumit Bhatia. 2025. Exploring the role of diversity in example selection for in-context learning. *Preprint*, arXiv:2505.01842.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *CoRR*, abs/2404.02060.

Yafu Li, Xuyang Hu, Xiaoye Qu, Linjie Li, and Yu Cheng. 2025. Test-time preference optimization: On-the-fly alignment via iterative textual feedback. *Preprint*, arXiv:2501.12895.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill 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, Dublin, Ireland and Online. Association for Computational Linguistics.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. 2023. Dr.icl: Demonstration-retrieved in-context learning. *Preprint*, arXiv:2305.14128.

Costas Mavromatis, Balasubramaniam Srinivasan, Zhengyuan Shen, Jiani Zhang, Huzefa Rangwala, Christos Faloutsos, and George Karypis. 2023. Which examples to annotate for in-context learning? towards effective and efficient selection. *Preprint*, arXiv:2310.20046.

MetaAI. 2024. Introducing meta llama 3: The most capable openly available llm to date.

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.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. *Preprint*, arXiv:2212.07677.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

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, Toronto, Canada. Association for Computational Linguistics.

Zhe Xu, Jiasheng Ye, Xiaoran Liu, Xiangyang Liu, Tianxiang Sun, Zhigeng Liu, Qipeng Guo, Linlin Li, Qun Liu, Xuanjing Huang, and Xipeng Qiu. 2025. Detectiveqa: Evaluating long-context reasoning on detective novels. *Preprint*, arXiv:2409.02465.

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*.

Mo Yu, Tsz Ting Chung, Chulun Zhou, Tong Li, Rui Lu, Jiangnan Li, Liyan Xu, Haoshu Lu, Ning Zhang, Jing Li, and Jie Zhou. 2025. Prelude: A benchmark designed to require global comprehension and reasoning over long contexts. *Preprint*, arXiv:2508.09848.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

A Correlating Factor Influencing In-Context CoT Success

Experiment Settings. The experiments were conducted on two datasets (i.e., BANKING77 and GSM8K), both exhibit an overall trend of improvement as the number of in-context examples increases, despite some fluctuations in GSM8K. The evaluation is performed on the two smaller LLMs (LLaMA 3.1, Qwen 2.5 (7B)). From the test set, the first 300 instances are retrieved to make a comparable evaluation to the testing accuracy. In each retrieved instance, the latter half of its question is masked and an external LLM, `LLaMA-3.3-70B-Instruct` is used to generate continuations for the masked questions to evaluate task understanding. The MCQA tasks are constructed under two distinct conditions:

1. Task A: LLaMA 3.3 is instructed to avoid the topic of the original task and generate each continuations in entirely different domains. This is to avoid LLM from locating the ground-truth continuation easily with only the option information. The clear differentiation between the domains of options indicates the task understanding. With more in-context examples, the accuracy of Task A is expected to continuously increase or plateau after a certain threshold. Examples of Task A are shown in Appendix D.1.

2. Task B: LLaMA 3.3 is allowed to generate continuations without the above restrictions, leading to highly semantically similar option sets with different expressions. Computing the averaged pairwise cosine similarity among the 4 options, Task B has a mean similarity score of 0.726, which is significantly higher than Task A of 0.471. Examples of Task B are shown in Appendix D.2.

For human evaluation, a university graduate student is invited to answer 120 sampled questions provided with the in-context examples from the training set. Humans can easily infer the task objectives and identify the correct option by aligning it with the task domain of the training examples. Results showed that humans achieved 93.33% accuracy in Task A, significantly higher than their performance of 46.67% accuracy in Task B.

Result. In the BANKING77 dataset, we observed a clear positive trend in LLM performance as the number of in-context examples increased. When provided with in-context CoT with {question, CoT, answer}, LLMs are prompted to predict the correct continuation of masked questions. Surprisingly, as shown in Table 1, Task B consistently exhibited a higher spearman rank correlation with task accuracy than Task A across all LLMs. This trend contrasts with human performance.

To validate these findings in the reasoning task, we conducted the same experiment on the GSM8K dataset, creating analogous MCQA tasks under the same conditions. Same conclusion is drawn with Task B demonstrating a significantly stronger correlation with task accuracy compared to Task A. All the correlations with Task B are statistically significant, with a p-value smaller than 0.05. It indicates in-context CoT correlates more to data distribution than task understanding.

		Task	Correlation	p-value
BANKING77	LLaMA	A	0.3376	0.4135
		B	**0.8752**	**0.004**
	Qwen	A	0.6412	0.0867
		B	**0.9140**	**0.001**
GSM8K	LLaMA	A	0.6190	0.1017
		B	**0.7075**	**0.0496**
	Qwen	A	0.6337	0.0916
		B	**0.7933**	**0.0188**

Table 1: Correlation between task accuracy and accuracy of the two constructed MCQA tasks.

B Prompt formatting and LLM performance for each task

B.1 SuperGlue

We evaluate the Winograd Schema Challenge (WSC) for coreference resolution, and the Choice of Plausible Alternatives (COPA) for open-domain commonsense causal reasoning. Both are formatted as a binary-label classification task. The prompt for inference is presented in Figure 9 and 12, while the evaluation result is shown in Figure 10 and 11 respectively.

B.2 TREC

We evaluate the Text REtrieval Conference (TREC) Question Classification dataset with 50 fine class labels. The prompt for inference is presented in Figure 13, while the evaluation result is shown in Figure 14.

Given a query, answer yes or no to the query.

The predicted answer must come from the demonstration examples with the exact format. The examples are as follows:

Question: In the sentence "{text_1}", does the pronoun '{span2_text_1}' refer to {span1_text_1}?
Answer: {answer_1}
...
Question: In the sentence "{text_n}", does the pronoun '{span2_text_n}' refer to {span1_text_n}?
Answer: {answer_n}

Now predict the answer for the following query:

Question: In the sentence "{text_i}", does the pronoun '{span2_text_i'}' refer to {span1_text_i'}?

reply in the following format:
'Answer: [yes | no]'

Figure 9: Prompt for WSC task

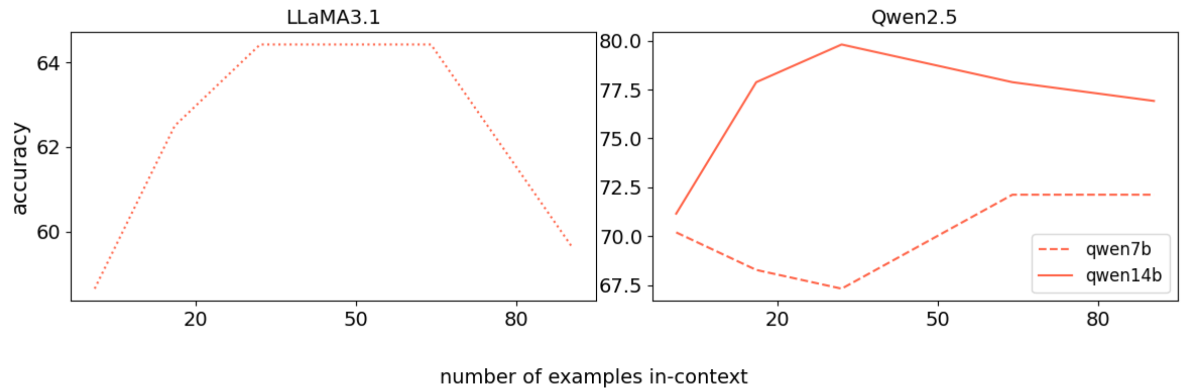


Figure 10: Performance on WSC

B.3 BANKING77

We evaluate the BANKING77 dataset with 77 fine-grained intents in the banking domain. The prompt for inference is presented in Figure 15, while the evaluation result is shown in Figure 16.

B.4 NLU

We evaluate the NLU dataset with 68 fine-grained intents in the conversational domain. The prompt for inference is presented in Figure 17, while the evaluation result is shown in Figure 18.

B.5 GSM8K

We evaluate the GSM8K dataset for grade school math word problems. The prompt for inference is presented in Figure 19, while the evaluation result is shown in Figure 20.

B.6 MATH

We evaluate the Mathematics Aptitude Test of Heuristics (MATH) dataset for mathematics competition problems, including the question types of counting_and_probability, prealgebra, geometry, precalculus, number_theory and algebra. The prompt for inference is presented in Figure 21, while the evaluation result is shown in Figure 22, 23, 24, 25, 26 and 27.

C Prompt for constructing MCQA Task A and B

C.1 Task A

The prompt to `LLaMA-3.3-70B-Instruct` for creating Task A is shown in Figure 28.

C.2 Task B

The prompt to `LLaMA-3.3-70B-Instruct` for creating Task A is shown in Figure 29.

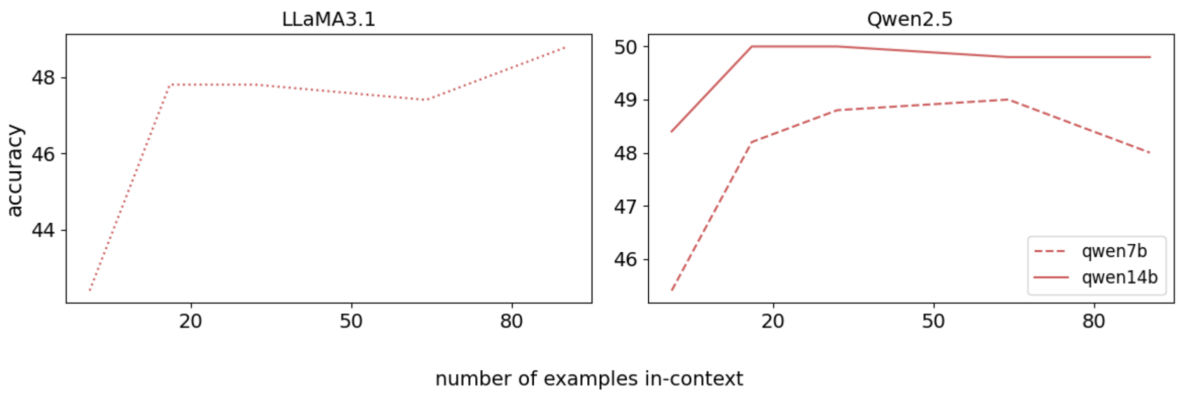


Figure 11: Performance on COPA

D Examples illustration of MCQA Task A and B

D.1 Task A

Two example illustrations in Task A, constructed for BANKING77 and GSM8K, are shown in Figure 30. The option highlighted in bold is the correct continuation of the incomplete question (Q).

D.2 Task B

Two example illustrations in Task B, constructed for BANKING77 and GSM8K, are shown in Figure 31. The option highlighted in bold is the correct continuation of the incomplete question (Q).

E Prompt formatting for Task A and B

The unified prompt for inference is presented in Figure 32.

Answer in A or B.

The predicted answer must come from the demonstration examples with the exact format. The examples are as follows:

Premise: {premise$_1$}
Question: What is the {question$_1$} for this?
Options:
A. {choice1$_1$}
B. {choice2$_1$}
Answer: {answer$_1$}

...

Premise: {premise$_n$}
Question: What is the {question$_n$} for this?
Options:
A. {choice1$_n$}
B. {choice2$_n$}
Answer: {answer$_n$}

Now predict the answer for the following query:

Premise: {premise$_i$}
Question: What is the {question$_i$} for this?
Options:
A. {choice1$_i$}
B. {choice2$_i$}

reply in the following format:
'Answer: [A | B]'

Figure 12: Prompt for COPA task

Given a question, predict the label of the question. You can only make predictions from the following categories: {LIST_OF_CATEGORIES}
Please predict the label of the FINAL question with the provided demonstration example queries as follows:

question: {question$_1$}
label: {label$_1$}
...
question: {question$_n$}
label: {label$_n$}

Now predict the answer for the following query:

question: {question$_i$}

reply in the following format:
'label: [category_name]'

Figure 13: Prompt for TREC task

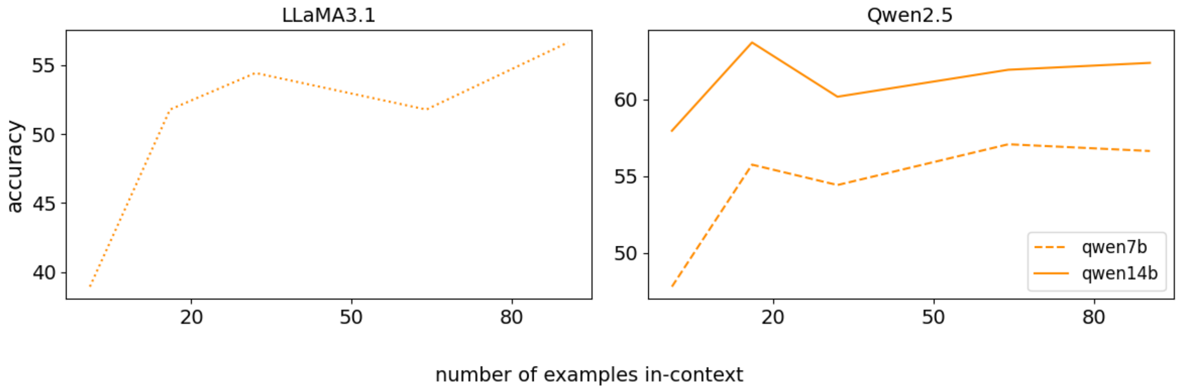


Figure 14: Performance on TREC

Given a question, predict the label of the question. You can only make predictions from the following categories: {LIST_OF_CATEGORIES}
Please predict the intent category of the FINAL query with the provided demonstration example queries as follows:

service query: {question_1}
intent category: {label_1}
...
service query: {question_n}
intent category: {label_n}

Now predict the intent category for the following query:

service query: {question_i}

reply in the following format:
'intent category: [category_name]'

Figure 15: Prompt for BANKING77 task

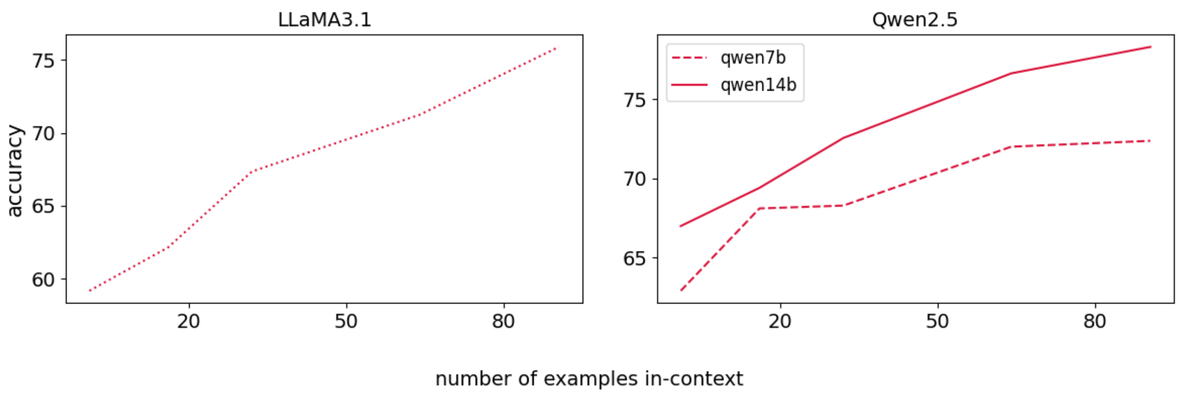


Figure 16: Performance on BANKING77

Given a question, predict the label of the question. You can only make predictions from the following categories: {LIST_OF_CATEGORIES}
Please predict the intent category of the FINAL utterance with the provided demonstration example queries as follows:

utterance: {question$_1$}
intent category: {label$_1$}
...
utterance: {question$_n$}
intent category: {label$_n$}

Now predict the intent category for the following utterance:

utterance: {question$_i$}

reply in the following format:
'intent category: [category_name]'

Figure 17: Prompt for NLU task

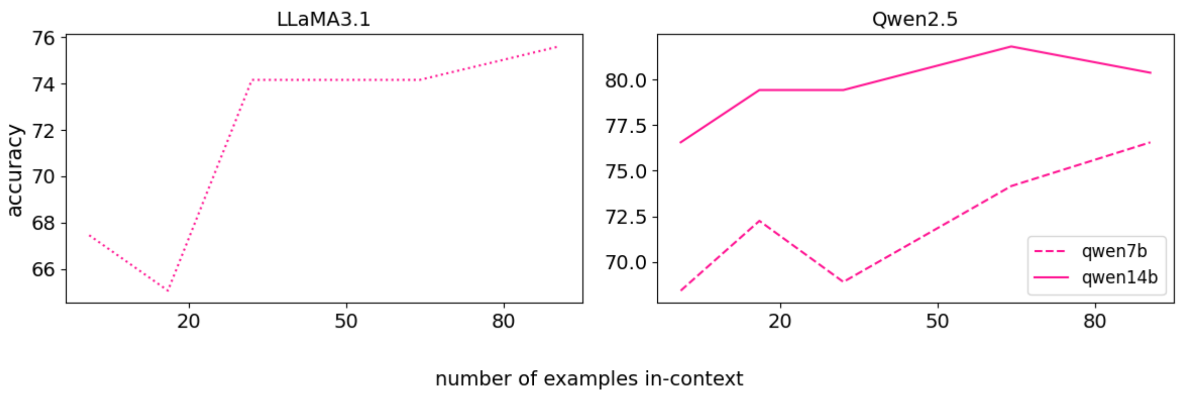


Figure 18: Performance on NLU

In the end of the response, add a summary 'The answer is [answer].'

Q: {question$_1$}
A: {CoT$_1$} {answer$_1$}
...
Q: {question$_n$}
A: {CoT$_n$} {answer$_n$}

Q: {question$_t$}
A: Let's think step by step.

Figure 19: Prompt for GSM8K task

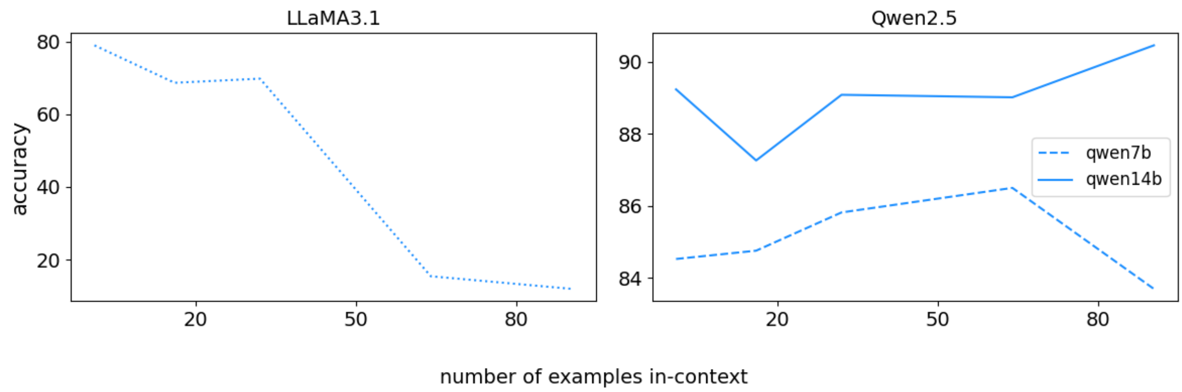


Figure 20: Performance on GSM8K

Write a response that appropriately completes the request and wrap the final answer inside \\boxed{ }.

Problem: {question$_1$}
Solution: {CoT_with_answer$_1$}
...
Problem: {question$_n$}
Solution: {CoT_with_answer$_n$}

Problem: {question$_t$}
Solution: Let's think step by step.

Figure 21: Unified prompt for MATH task

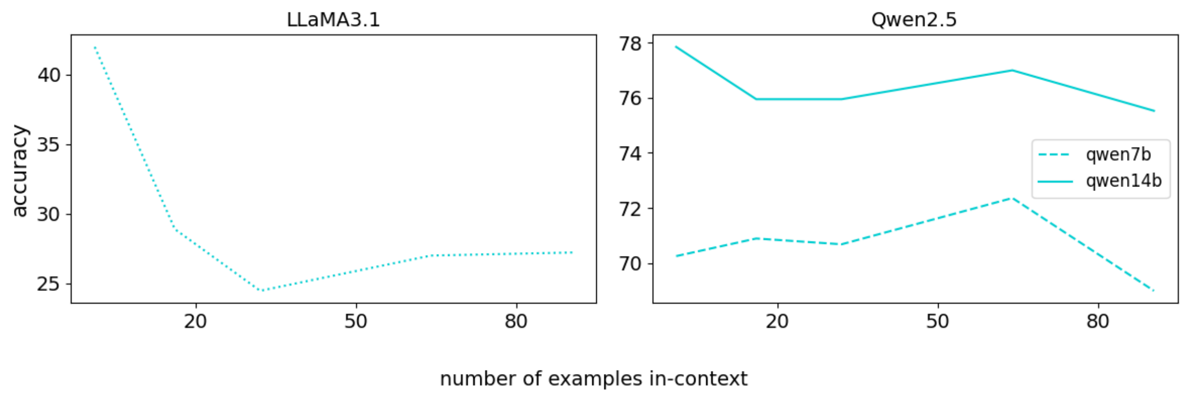


Figure 22: Performance on counting_and_probability

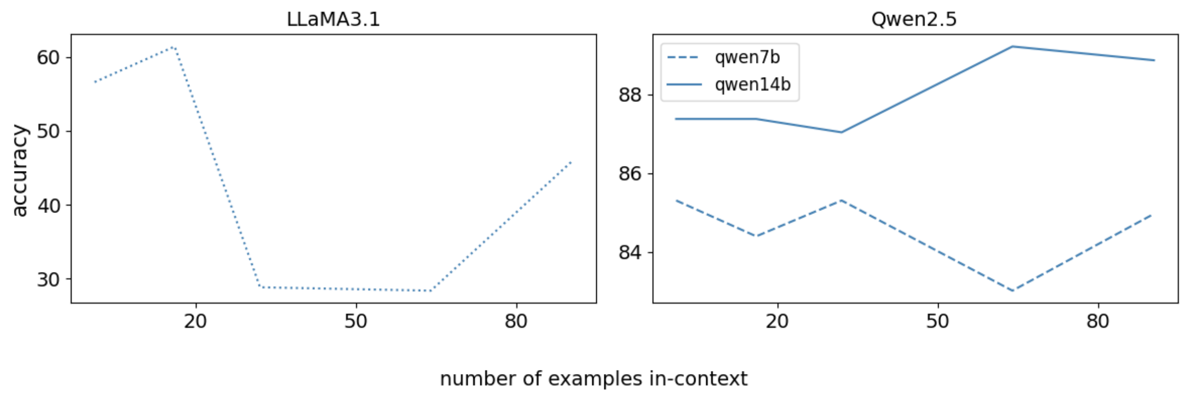


Figure 23: Performance on prealgebra

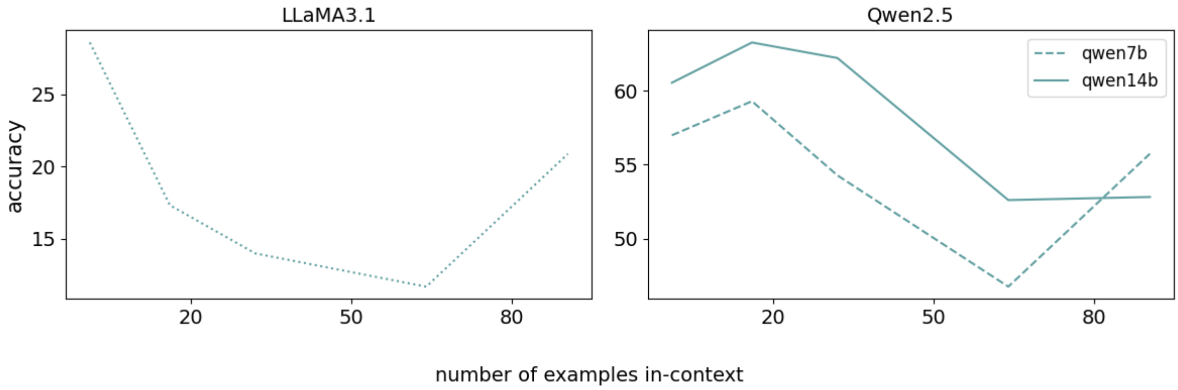


Figure 24: Performance on geometry

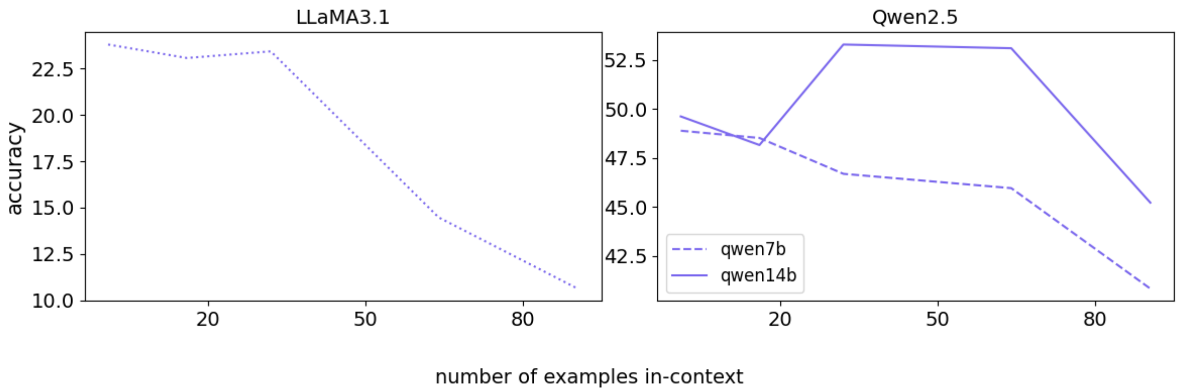


Figure 25: Performance on precalculus

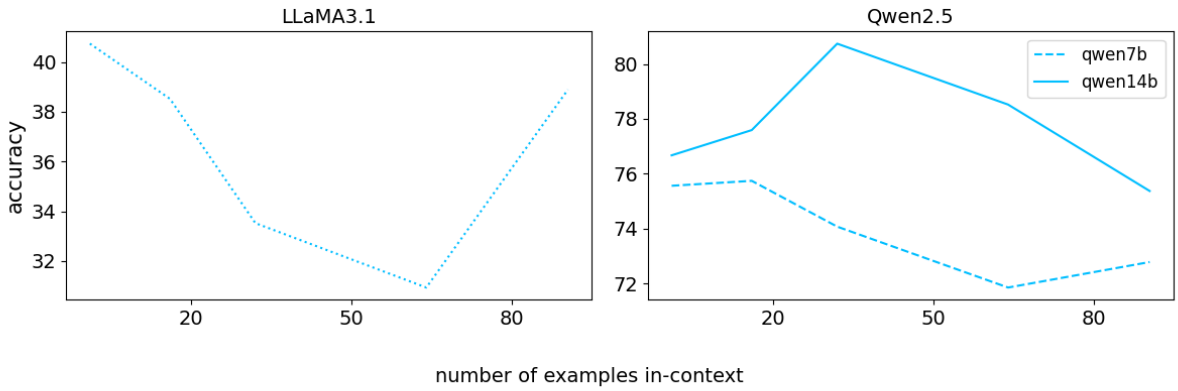


Figure 26: Performance on number_theory

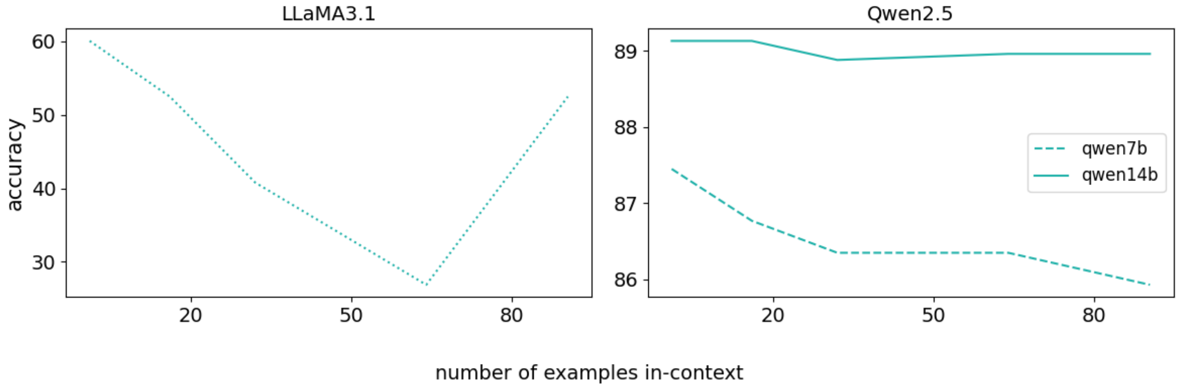


Figure 27: Performance on algebra

BANKING77:
Task: Task: Given an incomplete customer query, generate 6 unqiue and diverse continuations to complete the question. Ensure:
- The continuations avoid topics related to banking or customer service queries to ensure diversity.
- Each continuation explores different contexts or domains to reflect a variety of possibilities. Please diversify your generation with the provided example.

Input:
Q: {masked_question}

Output:
1. {groundtruth}
2.
3.
4.
5.

GSM8K:
Task: Given an incomplete question, generate 6 unique and diverse continuations to complete the question. Ensure:
- The continuations avoid topics related to math or grade school level difficulty to ensure diversity.
- Each continuation explores different contexts or domains to reflect a variety of possibilities. Please diversify your generation with the provided example.

Input:
Q: {masked_question}

Output:
1. {groundtruth}
2.
3.
4.
5.

Figure 28: Prompt for constructing Task A

Given an incomplete math question, generate a 6 version of continuation to complete the question. Ensure that the number of tokens in your completion is approximately equal to the number of tokens in the provided incomplete question.
Input:
Q: {masked_question}

Output:
1. {groundtruth}
2.
3.
4.
5.

Figure 29: Unified prompt for constructing Task B

BANKING77:
Q: When traveling, can I auto

Options:
A. track my daily expenses and stay within a set budget using a travel app?
B. navigate through unfamiliar cities using augmented reality maps?
C. unlock my hotel room door using a digital key on my smartwatch?
D. top-up my card at certain times?

GSM8K:
Q: Every day, Wendi feeds each of her chickens three cups of mixed chicken feed, containing seeds, mealworms and vegetables to help keep them healthy. She gives the chickens their feed in three separate meals. In the morning, she gives her flock of chickens 15

Options:
A. automated feeding system that dispenses the exact amount of feed at each meal, reducing waste and saving time. What are the potential advantages and disadvantages of using an automated feeding system for Wendi's flock?
B. varieties of vegetables, such as kale and carrots, to supplement their diet and provide essential nutrients. How can Wendi incorporate these vegetables into the chickens' feed in a way that is both cost-effective and efficient?
C. cups of feed. In the afternoon, she gives her chickens another 25 cups of feed. How many cups of feed does she need to give her chickens in the final meal of the day if the size of Wendi's flock is 20 chickens?
D. local farm-to-table initiative, where she sells the eggs produced by her chickens to nearby restaurants and cafes. What are some key marketing strategies that Wendi could use to promote her farm-to-table business and attract new customers?

Figure 30: Examples illustration of Task A. The correct answer is highlighted in bold.

BANKING77:
Q: If I request that my funds

Options:
A. be withdrawn, what are the minimum requirements?
B. be held, what currencies do you use?
C. be converted, what is the exchange rate?
D. be deposited, what is the maximum limit?

GSM8K:
Q: Toulouse has twice as many sheep as Charleston. Charleston has 4 times as many sheep as

Options:
A. Seattle. What is the combined total of sheep in Toulouse, Charleston, and Seattle if Seattle's sheep population is 10?
B. Seattle. How many sheep do Toulouse, Charleston, and Seattle have together if Seattle has 20 sheep?
C. Seattle. How many sheep are there in total in Seattle, Charleston, and Toulouse if Seattle has 15 sheep?
D. Seattle. If Seattle is home to 5 sheep, what is the total number of sheep in Charleston, Toulouse, and Seattle altogether?

Figure 31: Examples illustration of Task B. The correct answer is highlighted in bold.

BANKING77:
Your task is to choose the best option from the four provided that completes the question.
Do NOT solve or answer the question; ONLY respond with the correct option label (A, B, C, or D).

service query: {question$_1$}
intent category: {label$_1$}
...
service query: {question$_n$}
intent category: {label$_n$}

Final query: {mcqa_question}
A. {optionA}
B. {optionB}
C. {optionC}
D. {optionD}

GSM8K:
Your task is to choose the best option from the four provided that completes the question.
Do NOT solve or answer the question; ONLY respond with the correct option label (A, B, C, or D).

Question: {question$_1$}
Answer: {CoT$_1$} {answer$_1$}
...
Question: {question$_n$}
Answer: {CoT$_n$} {answer$_n$}

Final Question: {mcqa_question}
A. {optionA}
B. {optionB}
C. {optionC}
D. {optionD}

Figure 32: Unified prompt for Task A and B