On the Role of Chain of Thought in Long Test-Time 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, research has largely focused on non-reasoning 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 their understanding 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. Crucially, we find that demonstration selection and ordering remain critically important, while semantic similarity, which is a strong heuristic for few-shot ICL and RAG, becomes ineffective. We propose that effective manyshot CoT-ICL functions as a parameter-free, test-time learning process. Supporting this, we show that (1) self-generated demonstrations (where the model creates its own training curriculum) outperform ground-truth or stronger-model demonstrations, particularly for weaker models, and (2) smoothly ordered demonstrations (measured via embedding space curvature) significantly enhance performance, mirroring principles of curriculum learning. Our findings bridge manyshot ICL with test-time scaling paradigms, reframing the context window not as a static retrieval database, but as a dynamic, structured learning environment that triggers latent model capabilities.

1 Introduction

In-context learning (ICL) enables large language models (LLMs) to perform tasks by conditioning on a sequence of input-output demonstrations without updating their parameters (Min et al., 2022; von Oswald et al., 2023). Research has focused on improving ICL through strategies like selecting effective demonstrations (Sorensen et al., 2022; **?**). Recently, with the expansion of context windows, many-shot ICL has emerged, where dozens to hundreds of demonstrations can be provided, achieving performance competitive with fine-tuning (Agarwal et al., 2024; Bertsch et al., 2025). A consistent finding in this setting is that for non-reasoning tasks (e.g., classification), the impact of demonstration order diminishes with scale (Bertsch et al., 2025; Baek et al., 2025).

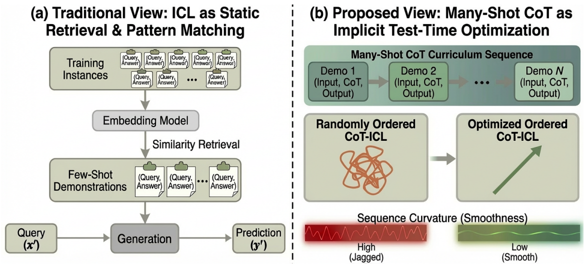


Figure 1: Reframing of CoT-ICL as test-time learning.

Meanwhile, chain-of-thought (CoT) prompting has become essential for complex reasoning tasks (e.g., arithmetic, narrative), where models generate intermediate reasoning steps before producing an answer (Wei et al., 2022; Kojima et al., 2022). Concurrently, the paradigm of test-time scaling investigates how to improve LLMs during inference without weight updates, through methods like sequential revision and parallel sampling (Snell et al., 2025; Lin et al., 2024). These threads connect naturally: many-shot ICL with CoT represents a fundamental form of test-time computation, where demonstrations guide the model's behavior and understanding.

However, a critical gap exists. Our understanding of many-shot dynamics derives almost entirely from studies of non-reasoning tasks. It remains unknown whether the established principles (e.g., that order matters less and similarity-based selection works) extend to many-shot CoT-ICL for reasoning. Does providing more reasoning demonstrations lead to reliable improvement, or does it introduce new instabilities? This question is practically important for deploying reasoning-capable LLMs and theoretically fundamental: it probes whether ICL for reasoning is merely large-scale pattern matching or a form of genuine in-context learning that follows pedagogical principles.

In this work, we demonstrate that the established rules of many-shot ICL break down for reasoning tasks. Through systematic experiments across model types (non-reasoning vs. reasoning-oriented) and tasks (non-reasoning vs. reasoning), our experiment uncover: (1) non-reasoning LLMs show negligible or negative gains from increasing CoT demonstrations; (2) Performance becomes more unstable with more demonstrations, indicating increased sensitivity to their order; and (3) Standard similarity-based retrieval consistently underperforms.

We explain these results by reframing effective many-shot CoT as in-context learning rather than pattern matching. We propose that successful demonstrations must be both understandable to the model and smoothly sequenced. We formalize this through two principles: (1) The Ease of Understanding: demonstrations should align with the model's current knowledge (explaining why self-generated demonstrations work best for weaker models); and (2) The Smoothness of Knowledge Progression: the conceptual transition between consecutive demonstrations should be gradual (quantifiable via the curvature of their embedding trajectory).

Building on these principles, we introduce Curvilinear Demonstration Selection (CDS), a practical method that orders demonstrations to minimize total conceptual curvature. This approach yields an average of 3.45% performance gains across both math and narrative reasoning tasks.

Our contributions are threefold: (1) We explore the dynamic with CoT-ICL; (2) We reframe effective many-shot CoT through the lens of comprehensibility and curricular smoothness, bridging ICL with insights from test-time learning; (3) We introduce and validate a practical, principle-driven method for demonstration ordering that advances many-shot reasoning as illustrated in Figure 1.

2 Related Works

2.1 Many-shot ICL

The extension of LLM context windows (Peng et al., 2024; Han et al., 2024) has enabled many-shot ICL, where models process significantly more demonstrations (Agarwal et al., 2024). Initial findings revealed that with sufficient demonstrations, model sensitivity to their ordering diminishes for standard classification tasks (Baek et al., 2025; Bertsch et al., 2025), suggesting a form of robustness with scaling. This led to a narrative that in many-shot settings, careful demonstration engineering may be unnecessary. However, these studies focused overwhelmingly on non-reasoning tasks (e.g., classification, simple QA) (Baek et al., 2025; Bertsch et al., 2025). Concurrent work on test-time scaling leveraging extended computation for self-improvement without parameter updates (Snell et al., 2024; Li et al., 2025), suggests that effective in-context learning can be viewed as a form of real-time optimization. Our work connects many-shot CoT-ICL to test-time learning, guided by two key principles that explain how learning occurs inside.

2.2 Chain-of-Thought

CoT prompting (Wei et al., 2022) decomposes reasoning into intermediate steps, substantially improving LLM performance on complex tasks. Subsequent studies like Tree-of-Thoughts (Yao et al., 2023) and Program-of-Thoughts (Chen et al., 2023) explore structured reasoning paths, while methods like rStar-Math (Guan et al., 2025) employ search algorithms for trajectory optimization. These approaches primarily focus on enhancing the reasoning process for a single query. In the ICL setting, Dr.ICL (Luo et al., 2023) demonstrates that retrieving relevant CoT demonstrations boosts few-shot performance. However, a critical gap remains with all existing CoT-ICL work operates in the few-shot settings. The fundamental question of how CoT demonstrations scale with context length and whether the principles of effective demonstration design change from few-shot to many-shot is largely unexplored. Our work positions many-shot CoT not merely as "more examples", but as a potential in-context curriculum that requires principled sequencing.

2.3 Demonstration Selection

Demonstration selection has long been studied for effective few-shot ICL. The dominant paradigm is similarity-based retrieval, where demonstrations semantically closest to the test query are selected (Liu et al., 2022; Wu et al., 2023; Kapuriya et al., 2025). This approach implicitly frames ICL as a form of nearest-neighbor pattern matching. Interestingly, this paradigm finds a direct analogy in Retrieval-Augmented Generation (RAG), where relevant context chunks are retrieved via embedding similarity (Lewis et al., 2020). Our work challenges whether this conclusion extends to reasoning tasks. We hypothesize that for CoT-ICL, effective demonstration selection is less about retrieving semantically similar examples and more about constructing a smooth learning sequence that facilitates conceptual understanding, acting as a shift from "retrieval for matching" to "retrieval for learning".

3 Experimental Setup

We establish a comprehensive experimental framework to study the factors influencing many-shot Chain-of-Thought In-Context Learning (CoT-ICL). Our framework focuses on three core dimensions: **Tasks Type** (non-reasoning vs. reasoning), **LLMs Type** (non-reasoning vs. reasoning LLMs), and **ICL Configuration** (format and scale).

3.1 Tasks Studied

Previous studies in many-shot (Li et al., 2024; Bertsch et al., 2025) have focused primarily on non-reasoning tasks. We bridge this gap by evaluating a diverse set of benchmarks spanning both classification and reasoning domains. All tasks are formulated as open-ended text generation. The model's raw output is matched against the reference answer or label with extra match.

Non-Reasoning Tasks. These tasks require minimal intermediate reasoning. We include tasks with varying label-space complexity, including SuperGLUE (Wang et al., 2019) (narrow label space), NLU (nlu, 2021), TREC (Hovy et al., 2001), and BANKING77 (Casanueva et al., 2020) (large label space).

Reasoning Tasks. These tasks require logical deduction and math derivation. We focus on mathematical reasoning with GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021). Additionally, we include DetectiveQA (Xu et al., 2025) for narrative reasoning over long contexts. All datasets provide ground-truth reasoning chains (C_i) for deriving the answer (y_i), enabling CoT-ICL.

3.2 LLMs Studied

We compare performance of varies LLMs in long context settings, categorized by their inherent reasoning design.

Non-Reasoning LLMs. These models lack a dedicated internal reasoning mechanism and are typically instructed-tuned for direct response, including 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`) and Qwen 2.5 (14B) (`Qwen2.5-14B-Instruct`).

Reasoning LLMs. These models are explicitly trained with a reasoning token (e.g., <think>), including Qwen 3 (8B) (`Qwen3-8B`), Qwen 3 (14B) (`Qwen3-14B`) (Yang et al., 2025) QwQ (32B) (`QwQ-32B`) (Qwen et al., 2025) and R1 (685B) (`DeepSeek-R1`) (DeepSeek-AI et al., 2025). For these models, we enable thinking token generation during inference.

To support the extended context required for many-shot evaluations (up to 131K tokens for Qwen family models), we applied official RoPE scaling configuration modifications.

3.3 ICL Configuration

We study performance from few-shot to many-shot under two prompting paradigms.

Traditional ICL. LLMs receives input-output pairs (x_i, y_i). Given an input x', it generates an answer $y' = \texttt{LLM}(x'|\{(x_i, y_i)\}_{i=1}^{k})$.

CoT-ICL. LLMs receives triplets of input, reasoning chain, and output, i.e., (x_i, C_i, y_i)). Given an input x', it generates both a reasoning chain C' and final answer $y' = \texttt{LLM}(x'|\{(x_i, C_i, y_i)\}_{i=1}^{k})$.

Context Scaling. The token length of CoT-ICL is substantially larger than that of traditional ICL (e.g., geometry problems are about 30 times longer than BANKING77 examples). Therefore, while models can process hundreds to thousands of traditional ICL demos, the context window typically limits CoT-ICL to hundreds demonstrations. Our analysis focuses on this scaling range (up to 128 demonstrations), where we observe the most informative dynamics between model capability, task type, and demonstration count.

4 Properties of CoT-ICL

4.1 Scaling with Non-Reasoning LLMs

While recent work demonstrates that many-shot ICL yields consistent gains for non-reasoning tasks (Bertsch et al., 2025; Baek et al., 2025), we find this scaling behavior does not generalize to reasoning tasks with CoT prompts. Figure 2 reveals that non-reasoning tasks exhibit steady improvement with more demonstrations, whereas math reasoning performance fluctuates or declines with non-reasoning LLMs for most of the tasks (especially for math reasoning tasks).

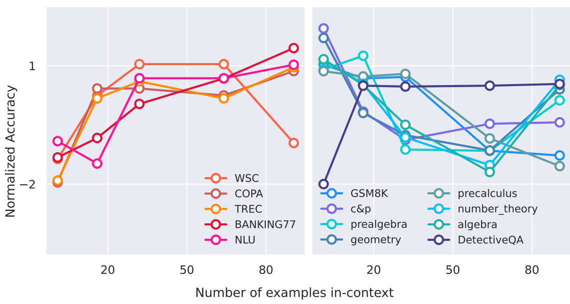
This failure to scale is not simply a limitation of model size. As shown in the left subplot of Figure 3, even the 70B-parameter Llama 3.3 shows negative gains. This contrasts with the effect of scaling observed in previous many-shot ICL, suggesting a qualitative difference in how LLMs process long CoT-ICL.

4.2 Scaling with Reasoning LLMs

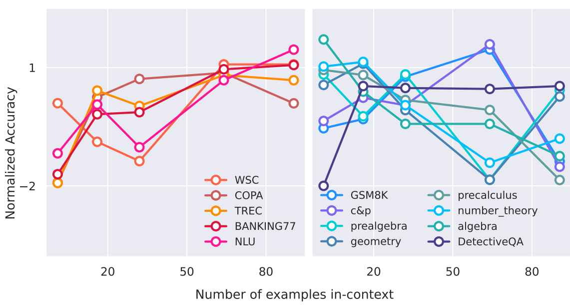
In contrast to non-reasoning LLMs, models with explicit reasoning capabilities exhibit a fundamentally different scaling pattern. As shown in right subplot of Figure 3, QwQ (32B) demonstrates clear positive scaling with additional demonstrations. This pattern holds for smaller reasoning-optimized models as well: the Qwen3 family (Figure 4) shows consistent performance gains as the number of demonstrations increases. LLMs with reasoning capabilities (enabled via thinking tokens or specialized training) successfully leverage additional CoT demonstrations to improve performance on reasoning tasks. The divergence in scaling behavior between model types highlights that the ability to benefit from many-shot CoT is not merely a function of pattern matching, but is intrinsically linked to a model's capacity for in-context reasoning.

4.3 Instability of Many-shot CoT-ICL

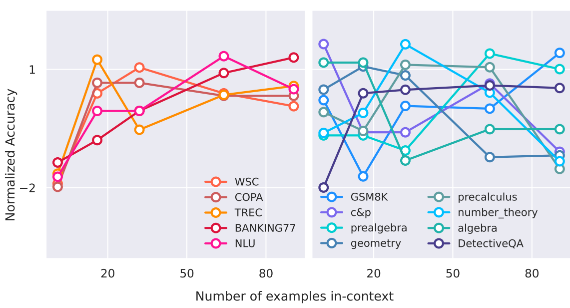
The divergent scaling patterns suggest that the sequence of demonstrations may be critical for CoT-ICL. To test this, we measure performance variance across five random orderings of the same demonstration set. Prior work finds that variance decreases with more demonstrations for non-reasoning tasks (Baek et al., 2025), indicating that order becomes less important. We observe the same for classification tasks in the left subplot in Figure 5.



(a) Llama-3.1-8B-Instruct



(b) Qwen2.5-7B-Instruct



(c) Qwen2.5-14B-Instruct

Figure 2: Scaling disparity between task types. Performance (normalized accuracy) of non-reasoning LLMs on classification tasks (warm colors) versus reasoning tasks (cool colors). The x-axis represents normalized accuracy (i.e., $\frac{x-\bar{x}}{\sigma_x}$ for accuracy x), while the y-axis indicates the number of in-context demonstrations.

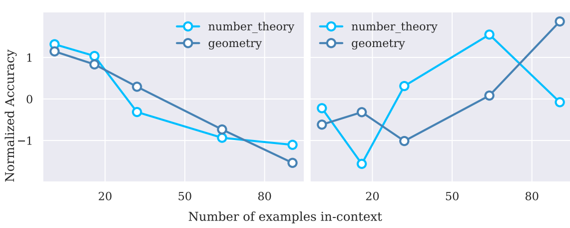


Figure 3: Scaling disparity between model types on math reasoning tasks. **Left:** Llama 3.3 (non-reasoning LLM) shows negative gains. **Right:** QwQ (32B) (reasoning LLM) shows clear positive scaling.

However, for reasoning tasks with CoT, we find the opposite trend. Variance increases with more demonstrations as shown in the right subplot in

Figure 4: Positive scaling of reasoning LLMs. The Qwen3 family (reasoning LLMs) demonstrates consistent performance improvements with more demonstrations on math reasoning tasks. **Left:** Qwen3 (8B) **Right:** Qwen3 (14B)

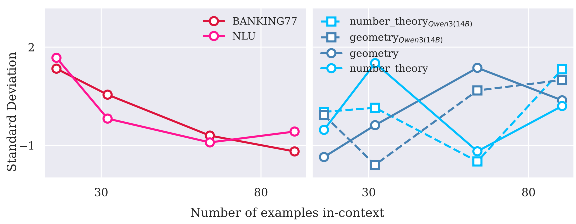


Figure 5: Standard deviation of performance across five random demonstration orders on classification tasks (warm colors) versus reasoning tasks (cool colors). Results shown for Qwen2.5 (14B) (**Left**, non-reasoning) and Qwen3 (14B) (**Right**, reasoning).

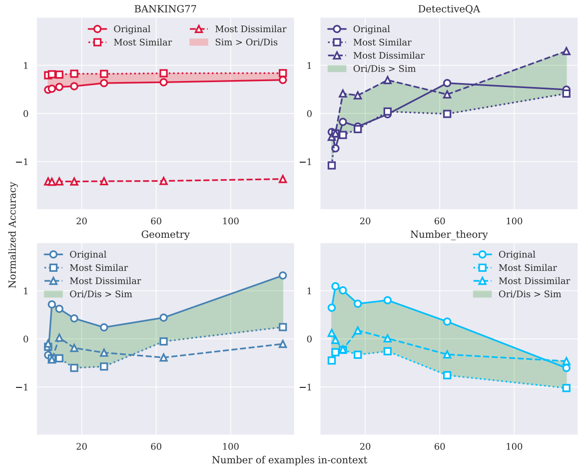


Figure 6: Performance with original(ori), similarity(sim) and dissimilar(dis) sets averaged across five LLMs. The area between the two sets is filled with colors, indicating the relative performance at each point. The normalization is performed with the mean and standard deviation computed over the concatenated sets of ori, sim and dis.

Figure 5. This holds for both non-reasoning and reasoning LLMs, revealing a key instability finding: Many-shot CoT exhibits increasing sensitivity to demonstration order as context length grows, unlike non-reasoning ICL where order becomes less important.

This increasing instability indicates that simply adding more CoT examples can introduce confusing signal. The model's performance becomes highly path-dependent, suggesting that the progression of reasoning steps across demonstrations is a critical, previously overlooked factor in CoT-ICL.

4.4 Rethinking the role of similarity

Given the importance of order, we investigate whether standard retrieval heuristics can identify helpful demonstrations. In few-shot ICL and retrieved-augmented generation (RAG), retrieving demonstrations semantically similar to the query is highly effective (Liu et al., 2022; Wu et al., 2023). We test this in the many-shot CoT setting using Qwen3-Embedding-4B (Zhang et al., 2025) to construct "most similar" and "most dissimilar" demonstration sets. 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. The training set instances form the candidate pool for constructing the similar and dissimilar example sets. Experiments are conducted and averaged over five LLMs, including Llama 3.1, Qwen 2.5 family (7B and 14B) and Qwen 3 family (8B and 14B).

The results in Figure 6 reveal an opposite pattern. For the BANKING77 classification task, similar examples outperform dissimilar ones, aligning to prior findings. For all three reasoning tasks (geometry, number_theory, DetectiveQA), the dissimilar set or the original set consistently outperform the similar set as the number of demonstrations increases. Performance for separated evaluated on reasoning and non-reasoning LLMs is in Appendix B and same conclusion is drawn when looking separately on different types of LLMs.

This failure suggests that reasoning tasks require a deeper, structural understanding of the problem space rather than surface-level pattern matching. Similar questions may cluster in solution strategy, providing redundant learning signal, while diverse (dissimilar) examples might better scaffold the model's understanding of the reasoning process itself.

5 Rethinking ICL: From Pattern Matching to Test-Time Learning

Our empirical findings in Section 4.3 and 4.4 present a puzzling contradiction, while CoT-ICL shows a promising trend with manyshot, it exhibits increasing sensitivity to demonstration order and fails to benefit from established heuristics like similarity-based selection. We propose a conceptual shift to explain these observations: rather than viewing ICL as mere pattern matching in an extended context, we should conceptualize it as test-time learning—a form of gradient-free optimization occurring within the forward pass. This reframing provides a principled basis for understanding what makes demonstrations effective and leads us to formulate two complementary principles for demonstration design.

5.1 Test-Time Learning

The success of simple heuristics like retrieving demonstrations similar to the query in such settings supports this pattern-matching view. However, our findings in Section 4.3, particularly the increasing variance with more demonstrations in reasoning tasks and the failure of similarity-based selection, directly contradict this interpretation for the many-shot CoT setting.

We posit that when provided with many demonstrations, especially those involving complex reasoning chains, the LLM engages in a more profound form of in-context learning: it is not merely recognizing a pattern, but actively constructing or refining an internal "algorithm" or reasoning schema based on the provided examples. This also aligns with the emergent perspective of test-time scaling and its effectiveness (Snell et al., 2024).

This learning analogy helps explain our key observations,

- **The importance of order:** Effective learning typically follows a curriculum—from simple to complex, or following logical progression. Random ordering disrupts this progression, leading to unstable "learning" outcomes.

- **The failure of similarity:** When learning a new concept, the most similar examples are often redundant and do not expand the model's understanding. Conversely, diverse examples that cover different facets of a problem can provide richer learning signals.

Effective demonstration selection for ICL requires retrieving pedagogically useful examples, those that facilitate learning the task itself rather than just providing answers. Similarly, test-time scaling methods like self-critique use multiple forward passes to iteratively refine outputs; many-shot ICL can be seen as a parallel, one-pass version of this refinement, where multiple demonstrations collectively shape the model's reasoning trajectory.

5.2 The Ease of Understanding

If ICL functions as test-time learning, then demonstrations must be comprehensible to the model within its current capabilities. Drawing from educational psychology, effective instruction operates within a learner's "zone of proximal development" (Benson, 2020), the space between what they can do independently and what they can achieve with guidance. We hypothesize that effective demonstrations must reside within the model's "zone of understandable reasoning."

Settings. To test this principle, we investigate whether demonstration efficacy depends more on reasoning quality or alignment with the model's own generative patterns. We generate CoT demonstrations by prompting each LLM on training instances and categorize them into three sets:

- **Correct Set (cr):** Model-generated CoT with correct final answers

- **Incorrect Set (wr):** Model-generated CoT with incorrect final answers

- **First Set (first):** The first generation for each instance, regardless of accuracy

These are compared against the dataset's groundtruth CoT (i.e., origin). Each LLM is prompted 10 times per training instance with temperature=1.0 to ensure diversity. Due to high accuracy on GSM8K, the wr set is constructed only for number_theory and geometry tasks. Additionally, we evaluate whether "better" CoT from stronger models (i.e., Qwen 2.5 (14B)) improves weaker model performance.

Results. Figure 7 reveals that the wr set (incorrect reasoning) consistently outperforms the original CoT and performs comparably to the cr set across both LLMs and tasks. This demonstrates that distributional alignment with the model's own reasoning style, even when flawed, contributes more to stable CoT prompting than the presence of correct answers.

Furthermore, self-generated CoT (any of cr, wr, or first sets) significantly mitigates the instability issues observed with origin CoT. The first set, the model's natural first attempt at each problem, also outperforms origin CoT (Figure 8), reinforcing that distributional alignment is paramount.

When using CoT from stronger models, we observe a mixed pattern: while occasional performance gains occur, instability persists (olive lines in Figures 7 and 8). This suggests that reasoning patterns too advanced for the target model can disrupt rather than enhance learning, similar to teaching advanced concepts before fundamentals.

Interpretation. These findings support that the effective demonstrations for in-context learning are those that the model can naturally comprehend and relate to its existing knowledge structures. Self-generated CoT, even when incorrect, provides such "understandable examples" by matching the model's own reasoning distribution, facilitating more stable test-time learning. It is also related to the LLM's internal ability to comprehend demonstration context. For demonstrations that are not understandable by LLMs (i.e., Qwen2.5 family and Qwen 3 (7B)), the benefit brings by self-generated CoT over groundtruth CoT is significant. But with a stronger LLM (i.e., Qwen 3 (14B)) that can understand well on the groundtruth CoT, this benefits shrink, as illustrated in Figure 8).

5.3 The Smoothness of Information Flow

Effective learning requires not just comprehensible individual examples, but a coherent progression between them. We hypothesize that smooth transitions between demonstrations facilitate the model's construction of a coherent reasoning schema, while abrupt conceptual jumps disrupt this process.

Quantifying Transition Smoothness To operationalize this principle, we conceptualize the sequence of demonstration embeddings as a trajectory through semantic space. We define the curvature between consecutive demonstrations as the angle between the vectors connecting them:

$$\theta_i = \arccos\left(\frac{(\mathbf{e}_i - \mathbf{e}_{i-1}) \cdot (\mathbf{e}_{i+1} - \mathbf{e}_i)}{\|\mathbf{e}_i - \mathbf{e}_{i-1}\|\|\mathbf{e}_{i+1} - \mathbf{e}_i\|}\right) \quad (1)$$

For an ordered sequence $O = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n]$ and their corresponding embedding $E = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\} \subset \mathbb{R}^d$, the total curvature $\Theta(E) = \sum_{i=2}^{n-1} \theta_i$, with lower values indicating smoother transitions between demonstrations. The detail algorithm is on Appendix A.

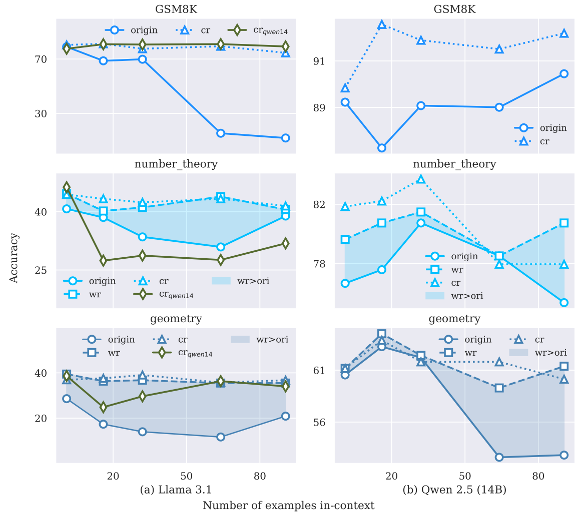


Figure 7: Performance of two sets of self-generated in-context CoT, including the set filtered with only correct answer(cr) and the set filtered with only wrong answer(wr). $\mathrm{cr}_{\mathrm{qwen14}}$ is prompting the LLaMA model with the in-context CoT generated by Qwen 2.5 (14B). **Left:** Llama 3.1 **Right:** Qwen 2.5 (14B)

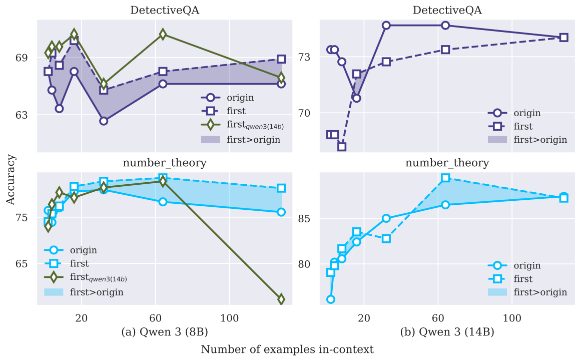


Figure 8: Performance of the first set of self-generated in-context CoT. $\mathrm{cr}_{\mathrm{qwen3(14b)}}$ is prompting the Qwen 3 (8B) model with the in-context CoT generated by Qwen 3 (14B). **Left:** Qwen 3 (8B) **Right:** Qwen 3 (14B)

Dimensionality reduction for efficiency. To facilitate efficient computation of smoothness and to capture both global and local structures, as well as linear and non-linear patterns, we project the embeddings into a lower-dimensional space using PCA (Maćkiewicz and Ratajczak, 1993) and UMAP (McInnes et al., 2018). Correlation is computed with 128 number of demonstration and we set the number of component $d' = 50$.

Projection Method: We compute a combined projection $P(E) \in \mathbb{R}^{n \times d'}$ as:

$$P(E) = \text{PCA}(E, d') + \text{UMAP}(E, d') \quad (2)$$

The PCA component $\text{PCA}(E, d')$ captures global linear structure, while the UMAP component $\text{UMAP}(E, d')$ preserves local nonlinear relationships crucial for reasoning tasks. The PCA projection is weighted by explained variance:

$$\text{PCA}(E, d') = E_{\text{pca}} \cdot \text{diag}(\sqrt{\boldsymbol{\lambda}}) \quad (3)$$

where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_{d'}]$ are the explained variance ratios.

Result. Our analysis reveals a negative correlation (r=-0.602) between ordering smoothness and performance across three math reasoning tasks. With the correlation of -0.654, -0.511 and -0.642 corresponding to geometry, number_theory and counting_and_probability tasks, it supports that smooth information flow facilitates effective in-context learning. Additionally, the finding that variance increases with more demonstrations in reasoning tasks in Section 4.3 can be reinterpreted as follow, With more demonstrations, the probability of encountering disruptive conceptual jumps increases, leading to greater outcome variability.

Pedagogical Analogy This principle mirrors effective textbook design: concepts are introduced progressively, with each chapter building smoothly upon the previous. Abrupt topic changes or missing prerequisites hinder learning. Similarly, in many-shot CoT-ICL, demonstrations must be ordered to create a "conceptual curriculum" that guides the model from basic to advanced reasoning steps.

6 Curvilinear Demonstration Selection

Based on the strong correlation between curvature and performance established in Section 5.3, we now introduce Curvilinear Demonstration Selection (CDS), a practical method for optimizing demonstration ordering in many-shot CoT-ICL. The core insight is that minimizing total curvature along the demonstration sequence corresponds to creating a smoother learning progression, analogous to how textbooks organize chapters by gradually increasing conceptual difficulty.

Method. Finding the global minimizer of Θ can computationally intractable for large n, but can be effectively approximated by formulating it as a

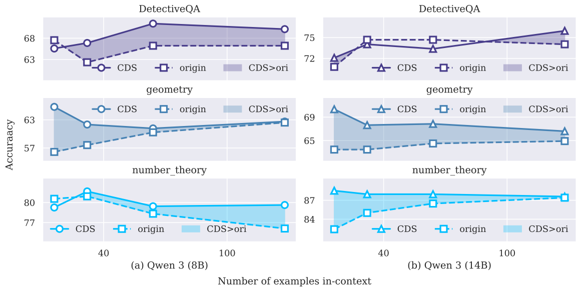


Figure 9: Performance comparison between CDS ordered CoT-ICL and originally ordered CoT-ICL. **Left:** Qwen 3 (8B) **Right:** Qwen 3 (14B)

Traveling Salesman Problem (TSP). We solve this TSP using a nearest neighbor heuristic with 2-opt optimization (Croes, 1958).

Dimensionality reduction. Since we cannot adopt $d' = 50$ for all number of demonstrations, we adopt the following strategy. For n demonstrations, we set the number of components d' as:

$$d' = \left\lfloor \frac{n}{5} \right\rfloor \times 5 \quad (4)$$

This rounding to the nearest multiple of five ensures computational efficiency while maintaining sufficient expressivity. For example, with $n = 128$, we use $d' = 125$ components.

Result. We evaluate CDS on three challenging reasoning tasks: geometry proof generation, number theory problem solving, and DetectiveQA logical reasoning. Figure 9 shows the performance comparison across different ordering strategies. Result shows that CDS outperform all baselines across the evaluated tasks, with average improvements of 3.45%.

7 Conclusion

We have shown that many-shot chain-of-thought in-context learning (CoT-ICL) does not follow the same property as standard many-shot ICL. To explain this, we reframe the ICL from pattern matching to in-context "learning", explained with two principles. Based on these, we propose CDS, a method that orders demonstrations to ensure smooth conceptual transitions. Our work reframes demonstration selection as a retrieval-for-learning problem. By designing demonstrations that teach rather than just match, we can build more robust and capable reasoning systems.

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. In addition, the effectiveness of CDS depends on the quality of the underlying embeddings. Though, Qwen3-Embedding-4B shows a promising performance on both narrative and math reasoning.

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*.

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.

Janette B Benson. 2020. *Encyclopedia of infant and early childhood development*. Elsevier.

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*.

Georges A Croes. 1958. A method for solving traveling-salesman problems. *Operations research*, 6(6):791–812.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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*.

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*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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.

Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

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 Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

A Algorithm for Curvature-Performance Correlation

To quantify the relationship between demonstration ordering smoothness and ICL performance, we develop Algorithm 1. The algorithm takes as input multiple orderings of demonstrations and their corresponding performance scores, and outputs a correlation coefficient between ordering smoothness and performance.

B Analysis of Similarity in Different LLM Types

Result in Figures 10 and 11 shows the performance comparison in different types of LLMs.

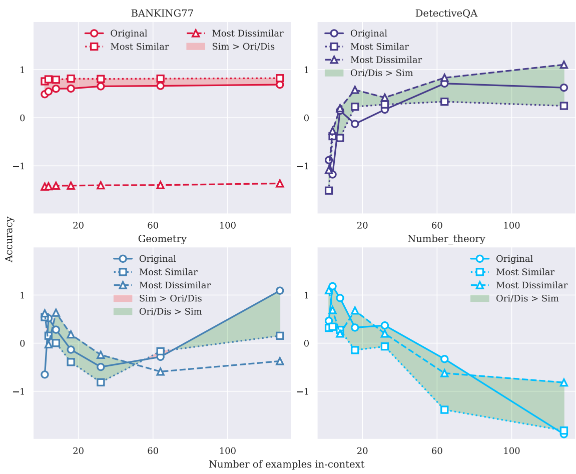


Figure 10: Performance with original (ori), similarity(sim) and dissimilar(dis) sets averaged across three non-reasoning LLMs. The area between the two sets is filled with colors, indicating the relative performance at each point.

C Prompt formatting and LLM performance for each task

C.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 12 and 13.

C.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 14.

Algorithm 1: Curvature-Performance Correlation Analysis

Input: For k different orderings:
- $E^{(1)}, E^{(2)}, \ldots, E^{(k)}$: embedding matrices where $E^{(j)} = [\mathbf{e}_1^{(j)}, \mathbf{e}_2^{(j)}, \ldots, \mathbf{e}_N^{(j)}]^\top$ represents the j-th ordering of N demonstration embeddings
- $S = [S_1, S_2, \ldots, S_k]$: performance scores for each ordering

Output: Correlation coefficient r between smoothness scores and performance scores

Initialize smoothness scores array
  $\mathbf{m} \leftarrow [0, 0, \ldots, 0]$ of length k

for *each dimensionality reduction method* $M \in \{PCA, UMAP\}$ **do**
  for $j = 1$ **to** k **do**
    # Step 1: Dimensionality reduction
    $\tilde{E}^{(j)} \leftarrow \text{ReduceDim}(E^{(j)}, M)$
    # Step 2: Compute curvature between consecutive demonstrations
    curvatures $\leftarrow [\,]$
    for $i = 2$ **to** $N - 1$ **do**
      $\mathbf{v}_1 \leftarrow \tilde{\mathbf{e}}_i^{(j)} - \tilde{\mathbf{e}}_{i-1}^{(j)}$
      $\mathbf{v}_2 \leftarrow \tilde{\mathbf{e}}_{i+1}^{(j)} - \tilde{\mathbf{e}}_i^{(j)}$
      if $\|\mathbf{v}_1\| > 0$ *and* $\|\mathbf{v}_2\| > 0$ **then**
        $\cos\theta_i \leftarrow \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}$
        $\theta_i \leftarrow \arccos(\cos\theta_i)$
        Append θ_i to curvatures
    # Step 3: Compute smoothness score
    $\bar{\theta}^{(j)} \leftarrow \text{mean}(\text{curvatures})$
    $\text{score}_M^{(j)} \leftarrow \frac{1}{1+\bar{\theta}^{(j)}}$
    # Step 4: Weighted combination (equal weighting for PCA and UMAP)
    $\mathbf{m}[j] \leftarrow \mathbf{m}[j] + 0.5 \times \text{score}_M^{(j)}$

 # Step 5: Compute correlation
$r \leftarrow \text{PearsonCorrelation}(\mathbf{m}, S)$
return r

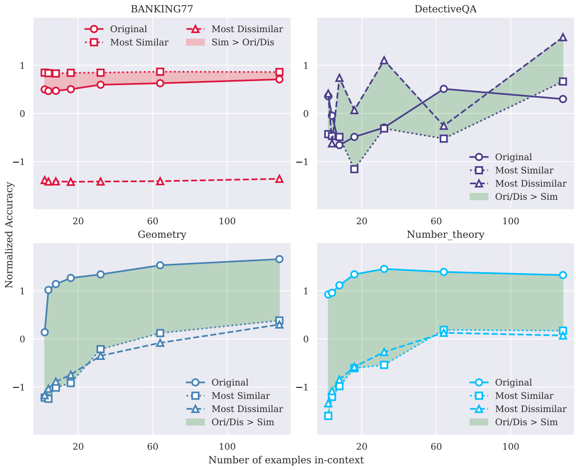


Figure 11: Performance with original (ori), similarity(sim) and dissimilar(dis) sets averaged across two reasoning LLMs. The area between the two sets is filled with colors, indicating the relative performance at each point.

C.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.

C.4 NLU

We evaluate the NLU dataset with 68 fine-grained intents in the conversational domain. The prompt for inference is presented in Figure 16.

C.5 GSM8K

We evaluate the GSM8K dataset for grade school math word problems. The prompt for inference is presented in Figure 17.

C.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 18.

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 12: Prompt for WSC task

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 13: 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 14: Prompt for TREC 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 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

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 16: Prompt for NLU task

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 17: Prompt for GSM8K task

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 18: Unified prompt for MATH task