

# Rethinking the Chain-of-Thought: The Roles of In-Context Learning and Pretrained Priors

Hao Yang<sup>1</sup>, Zhiyu Yang<sup>2</sup>, Yunjie Zhang<sup>3</sup>, Shanyi Zhu<sup>4</sup>, and Lin Yang<sup>1\*</sup>[0000-0001-9056-0500]

<sup>1</sup> School of Intelligence Science and Technology, National Key Laboratory for Novel Software Technology, Nanjing University

howyoung80@163.com, linyang@nju.edu.cn

<sup>2</sup> School of Computing and Information Systems, Singapore Management University  
kelvin.yangzhiyu@outlook.com

<sup>3</sup> Central South University

<sup>4</sup> School of Global Education and Development, International Chinese Language Education, University of Chinese Academy of Social Sciences  
2413589021@qq.com

**Abstract.** Chain-of-Thought reasoning has emerged as a pivotal methodology for enhancing model inference capabilities. Despite growing interest in Chain-of-Thought reasoning, its underlying mechanisms remain unclear. This paper explores the working mechanisms of Chain-of-Thought reasoning from the perspective of the dual relationship between in-context learning and pretrained priors. We first conduct a fine-grained lexical-level analysis of rationales to examine the model’s reasoning behavior. Then, by incrementally introducing noisy exemplars, we examine how the model balances pretrained priors against erroneous in-context information. Finally, we investigate whether prompt engineering can induce slow thinking in large language models. Our extensive experiments reveal three key findings: (1) The model not only quickly learns the reasoning structure at the lexical level but also grasps deeper logical reasoning patterns, yet it heavily relies on pretrained priors. (2) Providing sufficient exemplars shifts the model’s decision-making from pretrained priors to in-context signals, while misleading prompts introduce instability. (3) Long Chain-of-Thought prompting can induce the model to generate longer reasoning chains, thereby improving its performance on downstream tasks.

**Keywords:** Chain-of-Thought · Large Language Models · In-Context Learning · Pretrained Priors.

## 1 Introduction

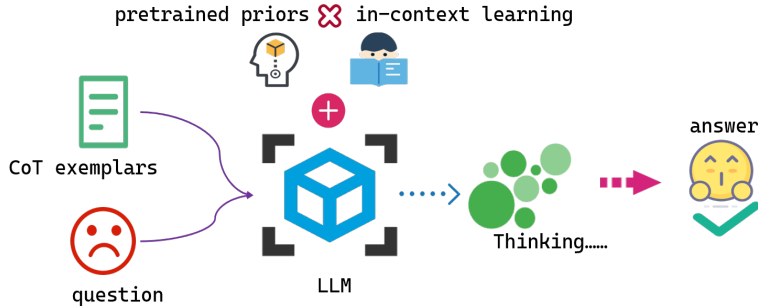
As models scale up, large language models (LLMs) exhibit emergent **In-Context Learning** (ICL) capabilities [2]. ICL enables LLMs to perform tasks by leveraging a few exemplars presented as demonstrations. Relying on ICL, **Chain-of-Thought** (CoT) prompting incorporates rationales into exemplars, guiding

---

\* Corresponding author

models to solve problems step by step and significantly improving performance across various downstream tasks [30]. The reasoning process in CoT prompting can be formalized as:

$$\text{Input: } X = \{(q_i, r_i, a_i)\}_{i=1}^N, q$$



**Fig. 1.** Framework highlighting the synergy between **Pretrained Priors** and **In-Context Learning** in powering CoT reasoning.

where  $(q_i, r_i, a_i)$  represents the  $i$ -th exemplar consisting of a question  $q_i$ , a rationale  $r_i$ , and an answer  $a_i$ ;  $N$  denotes the number of exemplars; and  $q$  is the target question. The model generates a response  $(r, a)$  by conditioning on the exemplars and the question:

$$P(r, a \mid X, q) = P(r \mid X, q) \cdot P(a \mid X, q, r)$$

CoT prompting has become an essential technique for enhancing model reasoning capabilities. Previous studies have extensively explored how CoT affects reasoning performance, focusing on factors such as difficulty, step length, the number of exemplars, and order [4,8,30,32].

However, the underlying mechanism of CoT prompting remains inconclusive. Inspired by [14,20], this paper explores the dual relationship between ICL and pretrained priors in CoT reasoning. In studies on the mechanism of ICL, [18] maintained that models primarily relied on pretrained priors during inference without acquiring new task-specific knowledge. This view was reinforced by [23,10], who argued that LLMs struggle to overcome their pretrained preferences. In contrast, [20,31] demonstrated that model scaling enables the learning of novel input-label mappings, although this capability emerged only in larger models (exceeding 66B parameters) rather than in smaller ones. Regarding CoT mechanisms, current research presents divergent perspectives: [22] proposed that co-variables in the prompts other than logical reasoning may be responsible for the performance improvements, while [16] attributed its effectiveness to task simplification achieved by decomposing complex problems. Furthermore, [15] suggested that models primarily imitated the CoT format through pattern recognition from

exemplars, and [17] provided theoretical analyses of how CoT amplifies models’ computational capabilities. Although these hypotheses require further empirical validation, they collectively advance our understanding of CoT’s fundamental mechanisms and offer valuable directions for future research.

As shown in Fig.1, we investigate the interaction between ICL and pretrained priors in CoT reasoning. Our study addresses three key questions: **(1) What do LLMs learn from CoT exemplars through ICL?** **(2) Can LLMs override pretrained priors through ICL?** **(3) Can prompt engineering leverage both ICL and pretrained knowledge to elicit slow thinking?** By disentangling the roles of ICL and pretrained priors, our findings shed light on the mechanisms behind reasoning emergence in LLMs, offering both theoretical insights into model cognition and practical implications for designing more effective prompts.

To answer these questions, we designed a series of experiments analyzing how LLMs integrate CoT exemplars and pretrained priors in decision-making. First, we provided models with both task-specific and task-agnostic exemplars to examine how variations in input influenced reasoning behavior. Through a fine-grained input-output analysis at the lexicon level, we quantified the reliance of LLMs on ICL signals versus pretrained priors. To assess robustness, we introduced controlled noise into exemplars and progressively increased the number of noisy samples to observe model behavior under misleading information. These previous experimental results lead to the third research question: given both ICL signals and pretrained priors, can prompt engineering guide models to generate longer CoT rationales? To test this, we applied prompts that encouraged slow thinking, aiming to generate extended CoT rationales and improve downstream performance. Our findings revealed three key insights:

- LLMs learn reasoning patterns from in-context exemplars, while CoT reasoning still remains influenced by pretrained semantic priors.
- Smaller models (8B) can map inputs to outputs via CoT prompting, with an increased number of exemplars shifting reliance from pretrained priors to ICL signals. Additionally, low-quality exemplars increase instability.
- Prompt engineering can induce longer CoT outputs, improving downstream performance and suggesting a path toward model self-evolution.

## 2 Experimental Setup

### 2.1 Evaluation Tasks

Given CoT’s pronounced impact on reasoning tasks, we focus on arithmetic, commonsense, and symbolic reasoning, following [4,8,16,30]. For arithmetic reasoning, we employed GSM8K [5] and MATH-500 [13], two widely adopted benchmarks in mathematical reasoning. For commonsense reasoning, we used Date Understanding [1], which features compositional questions designed to assess commonsense. For the symbolic reasoning task, we considered two tasks: the Coin Flip task [30] and the Last Letters Concatenation (four words) dataset [9].

The Coin Flip task assesses the ability to reason about probabilistic outcomes, while the Last Letters Concatenation dataset evaluates the capacity to manipulate and reason about symbolic representations.

## 2.2 Models & Prompts

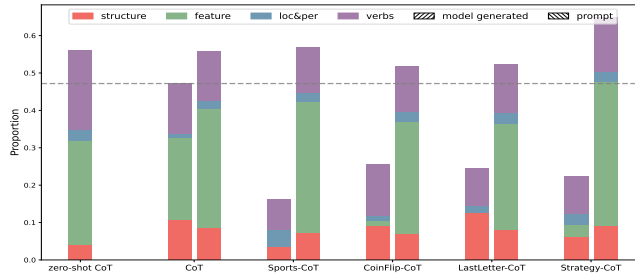
**Models.** We conducted experiments using publicly available pretrained models of various series and sizes, including Gemma2-9B, Gemma2-27B [25], LLaMA3.1-8B [26], and Qwen2.5-32B [27]. Following [30], we applied greedy decoding. Due to space constraints, only a subset of the results is presented in the main text.

**Prompts.** We employed standard and few-shot CoT exemplars from [30] and a zero-shot CoT instruction (“*Let’s think step by step.*”) from [9]. Each dataset is associated with a task-specific CoT prompt. When a CoT prompt designed for one task is applied to a different task, it is referred to as a task-agnostic prompt. For example, in Section 3.1, we utilized task-agnostic prompts (e.g., Sports-CoT, Date-CoT, etc.) on the GSM8K dataset.

## 3 Experiment

### 3.1 RQ1: What do LLMs Learn from CoT Exemplars Through ICL?

This study compares CoT reasoning acquired through pretraining with that learned from exemplars through ICL, revealing what models extract from exemplars and the characteristics of pretrained CoT reasoning. We evaluated base models of various scales and series, including Gemma2-9B, Gemma2-27B, and LLaMA3.1-8B.

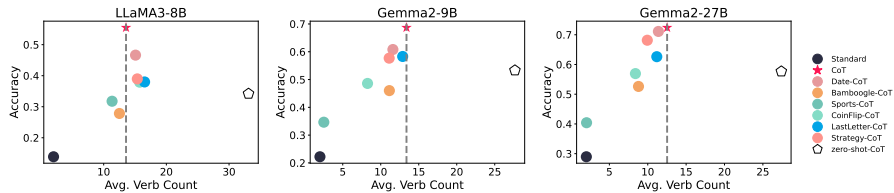


**Fig. 2.** Proportion analysis of rationale components: **Exemplars** (left) compare in-context **CoT** vs. **task-agnostic CoT** variants (Sports, Coin Flip, etc.); **Model-generated** content (right) contrasts **zero-shot CoT**, **CoT**, and **task-agnostic CoT** frameworks.

**Method.** From an ICL perspective, we conducted a fine-grained analysis of model-generated reasoning texts using the GSM8K test dataset. We decomposed exemplars into four key components: structure words, feature words, verbs, and location & person (loc&per) entities. Structure words capture textual flow (e.g.,

“The”, “So”, “Therefore”, “Then”, “Thus”). Feature words highlight task-specific elements such as numbers and operators (“+”, “-”, “\*”, “/”) in mathematical reasoning, extracted via string matching. Verbs represent reasoning actions, which we identified using the NLTK toolkit<sup>5</sup>. Loc&per words, also extracted with NLTK, denote entities related to location and individuals. We then explored model inference using varied prompts: zero-shot CoT, CoT, and task-agnostic CoT. The zero-shot CoT prompt reveals the model’s pretrained preference for CoT reasoning, while CoT and task-agnostic CoT prompts reflect the influence of in-context exemplars. This analysis provided insights into how different prompts shape reasoning behavior.

**Analysis.** Fig. 2 highlights distinct differences in inference content across the three prompt types. First, even with task-agnostic CoT, the model continued to perform mathematical reasoning based on the specific question. The notable rise in mathematical feature words indicates that task-agnostic CoT prompts did not alter the underlying reasoning behavior; LLMs remained guided by pretrained priors. Second, the CoT and task-agnostic CoT exemplars led to a significant increase in structural vocabulary compared to zero-shot CoT, suggesting that the model readily adopted lexical structures and mimicked exemplar language patterns. Finally, zero-shot CoT yielded a higher verb count than CoT. Under CoT and task-agnostic CoT prompts, verb usage declined. Given that verb frequency reflects sentence dynamism and causal expression [6,7,24], this finding suggests a deeper-level imitation of exemplar reasoning forms. We further analyzed verb usage to explore this effect. Building on [8,11,32], we note that reasoning abil-



**Fig. 3.** Scatter plot illustrating the relationship between the average number of **reasoning verbs** in model-generated content and **accuracy** under **CoT** and **task-agnostic CoT** settings.

ity is strongly influenced by the number of reasoning steps, which are typically considered at the sentence level. In this work, we explored the lexical level, hypothesizing that the number of reasoning actions also affects performance, with an optimal count potentially existing. Using the NLTK toolkit, we extracted and counted verbs from CoT and task-agnostic CoT prompt generations, computing the average number of reasoning verbs per sample across the test dataset. We then analyzed the relationship between prediction accuracy and verb count.

<sup>5</sup> <https://www.nltk.org/>

Fig. 3 shows that even task-agnostic CoT prompts can induce question-specific CoT reasoning. Some task-agnostic CoT results approached CoT performance and significantly outperformed zero-shot CoT, indicating that reasoning structure, rather than content, is crucial. We also observed a positive correlation between performance and the average number of reasoning verbs, with an optimal count evident. Below this optimal point, performance improved as the verb count increased; beyond it, performance declined. Additionally, the model exhibited a baseline CoT reasoning ability from pretraining. The weaker performance observed with zero-shot CoT compared to manual CoT may stem from overly divergent reasoning, where excessive actions introduce noise. In contrast, manual CoT maintained balanced reasoning depth, allowing the model to solve problems more effectively with fewer verbs. This observation suggests that the model captures deeper reasoning structures from exemplars.

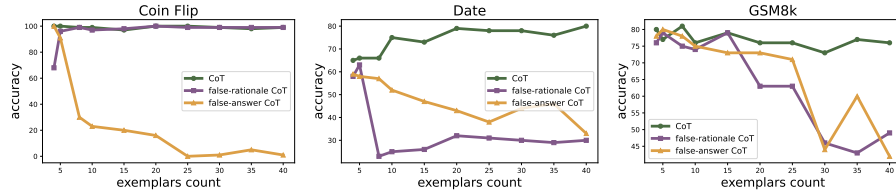
#### Takeaway

LLMs leverage ICL to capture fine-grained lexical structures and deeper reasoning patterns from exemplars, while pretrained priors continue to shape the reasoning process.

### 3.2 RQ2: Can LLMs Override Pretrained Priors Through ICL?

Prior studies [16,22,12] have shown that incorrect reasoning has little effect on LLMs performance. In this section, we examine whether the model’s confidence and accuracy remain stable as the number of exemplars increases. We treat the token generation probability at each time step as a measure of confidence and analyze how this confidence shifts with incorrect reasoning, as well as how accuracy evolves with an increasing number of incorrect exemplars.

**Method.** We adapted methods from [16,18,22] to create two contrastive prompt types. The first type, *false-answer CoT prompts*, retains the correct rationale but replaces the final answer. In the Coin Flip dataset, we inverted the answers by replacing “yes” with “no” and vice versa. In GSM8K, correct answers were replaced with random numbers, while in the Date dataset, correct dates were swapped with random alternatives. The second type, *false-rationale CoT prompts*, modifies the reasoning steps while keeping the final answer correct. In the Coin Flip dataset, key verbs such as “Flipped” and “is” were replaced with “not Flipped” and “not is”. For the Date dataset, dates were shifted by adding 30 days. In GSM8K, operators were swapped: “+” with “-”, “-” with “+”, “\*” with “/”, and “/” with “\*”. Due to the limited exemplars provided in [30], we used the small provided set and distilled 40 additional exemplars for each task from the training data using GPT-4 [19]. We then applied the above modifications to create the false-answer and false-rationale CoT prompts.

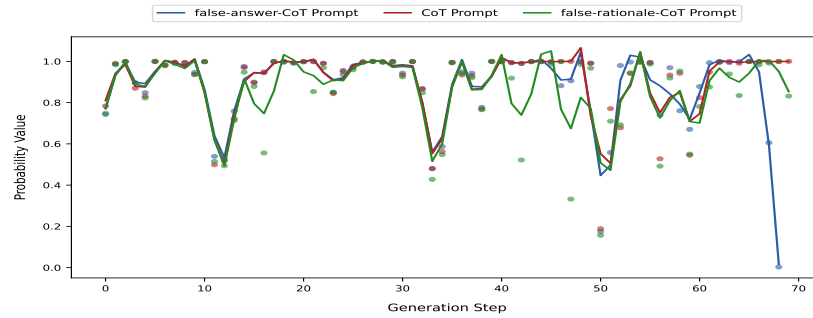


**Fig. 4.** Evolution of test accuracy as the number of **noisy** (false-answer, false-rationale) exemplars increases.

**Analysis.** As shown in Fig. 4, when provided with a small number of exemplars (e.g., 4-shot or 5-shot), the accuracy differences among CoT, false-rationale CoT, and false-answer CoT are minimal. This observation aligns with prior findings that noise in exemplars has little impact on model accuracy during reasoning [16,22]. However, as the number of exemplars increases, these differences become more pronounced.

We first examine the false-answer CoT prompts. In the Coin Flip task, accuracy declines most sharply as the number of false-answer exemplars increases. This is attributed to its closed-domain nature with a binary label space, where the model tends to learn an input-label mapping that leads to systematic label flipping. In contrast, for the Date and GSM8K tasks, which feature open-domain answer spaces, such mapping is less feasible, resulting in a more gradual decline in accuracy. Nevertheless, at 40-shot, accuracy drops by nearly half, highlighting the cumulative impact of noisy exemplars. Next, we analyze the false-rationale CoT prompts. In the Coin Flip task, accuracy remains relatively stable despite an increasing number of false-rationale exemplars, likely because the small label space allows the model to associate correct answers with certain patterns, minimizing accuracy fluctuations. However, in the Date and GSM8K tasks, noisy rationales significantly impair model reasoning as the exemplar count rises; at 40-shot, accuracy is halved, indicating severe degradation in reasoning ability under large-scale noise.

Our findings challenge previous claims. For instance, [31] suggested that smaller models cannot learn to flip labels; yet our results show that 8B model exhibit label flipping in closed-domain tasks when using CoT reasoning. Similarly, [22] argued that noisy rationales have little effect on reasoning accuracy, we demonstrate that accuracy deteriorates significantly with an increasing number of noisy exemplars. Overall, our analysis reveals an interaction between pre-trained priors and ICL in CoT reasoning. In early-shot settings, pre-trained priors dominate, stabilizing performance despite noise. As the number of exemplars increases, ICL signals strengthen and shift model decisions based on the provided examples. In closed-domain tasks, this shift leads to systematic label flipping, whereas in open-domain tasks, accuracy declines more gradually yet substantially. These findings underscore the need to balance exemplar quality and quantity to mitigate ICL-induced biases.



**Fig. 5. Probabilities evolution** of model-generated outputs under greedy decoding for three prompt types: **CoT** prompt, **false-answer CoT** prompt, and **false-rationale CoT** prompt.

**Case Study.** We further investigated the model’s internal mechanisms to understand how false-answer and false-rationale CoT prompts affect reasoning. Under greedy decoding, we recorded the token generation probabilities at each time step to reflect the model’s confidence. Figure 5 presents these probabilities over time step for the three prompt types: CoT, false-answer CoT, and false-rationale CoT. The raw probabilities (displayed as scatter points) were obtained by normalizing logits at each generation step, while smoothed curves—derived via Savitzky–Golay filtering<sup>6</sup>—highlight the overall trends. CoT prompts maintained stable probability values, indicating high confidence. In contrast, false-answer and false-rationale CoT prompts exhibited greater fluctuations, reflecting reduced confidence. This variability suggests that the model detected incorrect reasoning or misleading information, leading to instability. These findings emphasize the role of correct reasoning in reinforcing confidence: CoT prompts provide logical steps that bolster predictions, whereas misleading prompts introduce conflict and uncertainty. From an ICL perspective, correct reasoning enables the model to leverage contextual signals effectively, while misleading prompts undermine stability by contradicting pretrained knowledge. This underscores the importance of well-designed prompts for reliable reasoning.

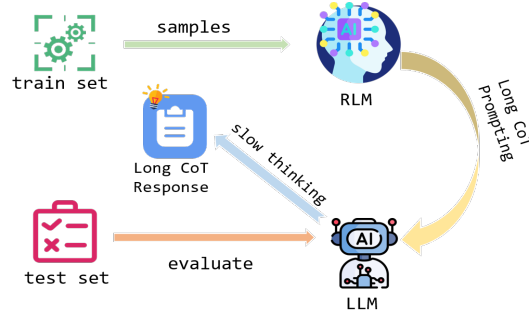
#### Takeaway

Increasing the number of exemplars shifts decision-making from pre-trained priors to ICL signals. However, misleading prompts introduce instability, underscoring the importance of high-quality exemplars.

<sup>6</sup> [https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.savgol\\_filter.html](https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.savgol_filter.html)



### 3.3 RQ3: Can prompt engineering leverage ICL ability and pretrained knowledge to elicit slow thinking?



**Fig. 6.** Framework for employing **prompt engineering** to encourage **slow thinking** in LLMs.

Researchers enhance LLMs reasoning abilities by extending the length of CoT, leading to the development of Reasoning Language Models (RLMs). RLMs generate a large number of tokens during inference, a process known as *slow thinking* or *test-time scaling*. The results of two previous experiments indicate that LLMs effectively leverage ICL to adopt the reasoning styles of the exemplars while also utilizing pretrained knowledge. In this section, we investigate whether LLMs can rely on ICL and pretrained knowledge to engage in slow thinking, thereby improving their reasoning performance. As illustrated in Fig. 6, we distilled long CoT prompts from three open-source RLMs: DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-32B, and QwQ-32B. We then used the long CoT distilled as prompts to guide LLMs (LLaMA3.1-8B, Qwen2.5-32B) in reasoning, examining whether LLMs can learn the slow thinking reasoning approach through prompting. Specifically, we conducted tests on four datasets: GSM8K, MATH-500, Date, and Last Letter Concatenation. The experimental results are shown in Table 1.

By checking the generated content, we find that the model effectively emulates the reflective, retrospective, and summarization-based slow thinking of long CoT, leading to improved performance on downstream tasks. This effect is particularly pronounced in instruct models due to their strong instruction following capabilities. The data in the Table 1 indicates an optimal CoT reasoning length for the model, influenced by model capacity and task difficulty. This result is consistent with the findings of [32,33]. For instance, on GSM8K with the LLaMA3.1-8B-Instruct, performance declines when CoT length exceeds a certain point. In the 8B model, shorter CoT length yields strong performance, but longer length leads to degradation. Similarly, in the Qwen2.5-32B model, the optimal CoT length increases with model size, and shorter CoT lengths result in poorer reasoning performance.

**Table 1.** Accuracy and Avg. tokens total of different models with various CoT prompts on datasets

Model	Prompt	Accuracy (Avg. tokens total)			
		GSM8K	MATH-500	DATE	Last-Letter
LLaMA3.1-8B	Manual-Short-CoT	0.2676(92)	0.112(101)	0.6016(49)	0.5504(69)
	DS-LLaMA8B-Long-CoT	0.4571(364)↑	0.120(534)↑	0.6203(252)↑	0.7217(439)↑
	DS-Qwen32B-Long-CoT	0.4617(343)↑	0.125(575)↑	0.6149(277)↑	0.7020(484)↑
	QwQ-32B-Long-CoT	0.3449(650)↑	0.118(810)↑	0.6338(252)↑	0.7560(997)↑
LLaMA3.1-8B-Instruction	Manual-Short-CoT	0.7695(114)	0.264(274)	0.6504(52)	0.4173(69)
	DS-LLaMA8B-Long-CoT	0.8248(377)↑	0.308(765)↑	0.6991(280)↑	0.7600(438)↑
	DS-Qwen32B-Long-CoT	0.7901(383)↑	0.304(826)↑	0.6449(322)↓	0.8830(494)↑
	QwQ-32B-Long-CoT	0.6202(702)↓	0.298(767)↑	0.6856(368)↑	0.8487(589)↑
Qwen2.5-32B	Manual-Short-CoT	0.8104(150)	0.310(407)	0.6991(62)	0.7903(72)
	DS-LLaMA8B-Long-CoT	0.8195(425)↑	0.410(619)↑	0.7775(315)↑	0.8185(436)↑
	DS-Qwen32B-Long-CoT	0.8599(440)↑	0.392(698)↑	0.7667(368)↑	0.8302(472)↑
	QwQ-32B-Long-CoT	0.8777(511)↑	0.425(1082)↑	0.7639(427)↑	0.8161(631)↑
Qwen2.5-32B-Instruction	Manual-Short-CoT	0.8316(117)	0.590(415)	0.8292(108)	0.7802(192)
	DS-LLaMA8B-Long-CoT	0.8221(401)↓	0.602(688)↑	0.8886(379)↑	0.9052(1032)↑
	DS-Qwen32B-Long-CoT	0.8945(681)↑	0.626(586)↑	0.8721(344)↑	0.8649(1058)↑
	QwQ-32B-Long-CoT	0.8529(893)↑	0.638(1305)↑	0.8723(390)↑	0.8407(1105)↑

### Takeaway

Leveraging ICL and pretrained knowledge, LLMs can adopt slow thinking through prompt engineering, effectively generating long CoT reasoning.

## 4 Related Works

*In-Context Learning.* Recent studies have examined the role of labels in ICL. [35] revisit label randomization and report significant variance across tasks and models. [20] distinguish between label-independent and label-dependent learning by substituting labels with arbitrary tokens. [31] demonstrate that smaller models struggle with ICL when labels are replaced. However, these studies overlook probabilistic metrics, potentially underestimating the impact of label modifications. For instance, [20] suggest that the performance gap between random and default labels is insignificant for small models, while [31] argue that large models can override pretraining priors in context, whereas small models fail to adjust to flipped labels—claims that our findings in Section 3.2 challenge.

*Chain-of-Thought.* CoT developments have significantly improved performance, particularly in complex reasoning tasks such as arithmetic and commonsense reasoning [30]. This success has led to research on Least-to-Most prompting [36], and specialized techniques such as Program-of-Thought [3], Contrastive Chain-of-Thought [4]. Research has examined various factors influencing CoT to enhance prompt design. The use of diverse examples has been found to complement

performance [34]. Maintaining logical coherence, even in the presence of errors, retains performance [4,28,22,12]. However, understanding why CoT is effective remains limited. Some studies suggest that reasoning abilities arise during pre-training, with exemplars guiding generation [29,4,21], though explicit empirical support is lacking. Other research argues that intermediate steps serve as templates rather than aiding task-solving [32]. Further findings indicate that imitation improves style adherence but not factuality or problem-solving [15].

Unlike the above studies, our work explores CoT mechanisms from the perspectives of ICL and pretrained priors, highlighting their interplay for a cohesive understanding.

## 5 Conclusion

Our study demonstrates that LLMs, through ICL, not only capture lexicon-level reasoning structures from CoT exemplars but also internalize deeper reasoning logic. Providing sufficient exemplars shifts the model’s decision-making from pre-trained priors to ICL signals. Conversely, misleading prompts induce instability, highlighting the crucial role of exemplar quality. Moreover, by leveraging ICL and pretrained knowledge, prompt engineering enables models to partially emulate slow thinking, offering a promising path for self-evolution.

## 6 Limitation

This study explores the mechanism of CoT reasoning from the perspective of the interplay between ICL and pretrained priors, further examining the potential of inducing slow thinking through prompt engineering. However, our experiments are limited to a few datasets, mainly in mathematics, leaving more complex reasoning tasks to be tested. Additionally, the impact of different levels of noise on CoT reasoning requires further investigation. Finally, our use of greedy decoding for long CoT often resulted in endless repetitions, indicating the need to explore alternative model settings for better performance.

## References

1. BigBench: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (2023)
2. Brown, T., Mann, B., Ryder, N., et al.: Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
3. Chen, W., Ma, X., et al.: Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research* (2023)
4. Chia, Y.K., Chen, G., et al.: Contrastive chain-of-thought prompting (2023)
5. Cobbe, K., Kosaraju, V., et al.: Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021)

6. Graesser, A.C., Singer, M., Trabasso, T.: Constructing inferences during narrative text comprehension. *Psychological review* **101**(3), 371 (1994)
7. Hopper, P.J., Thompson, S.A.: Transitivity in grammar and discourse. *language* **56**(2), 251–299 (1980)
8. Jin, M., Yu, Q., et al.: The impact of reasoning step length on large language models. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 1830–1842. Association for Computational Linguistics (2024)
9. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: Advances in Neural Information Processing Systems. vol. 35, pp. 22199–22213. Curran Associates, Inc. (2022)
10. Kossen, J., Gal, Y., Rainforth, T.: In-context learning learns label relationships but is not conventional learning. In: International Conference on Learning Representations (ICLR) (2024)
11. Lee, A., Che, E., Peng, T.: How well do llms compress their own chain-of-thought? a token complexity approach (2025), <https://arxiv.org/abs/2503.01141>
12. Li, D., Cao, S., et al.: Llms can easily learn to reason from demonstrations structure, not content, is what matters! arXiv preprint arXiv:2502.07374 (2025)
13. Lightman, H., Kosaraju, V., , et al.: Let’s verify step by step. arXiv preprint arXiv:2305.20050 (2023)
14. Lin, Z., Lee, K.: Dual operating modes of in-context learning. In: Forty-first International Conference on Machine Learning (2024)
15. Madaan, A., Yazdanbakhsh, A.: Text and patterns: For effective chain of thought, it takes two to tango (2022), <https://arxiv.org/abs/2209.07686>
16. Madaan, A., et al.: What makes chain-of-thought prompting effective? a counterfactual study. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 1448–1535. Association for Computational Linguistics (2023)
17. Merrill, W., Sabharwal, A.: The expressive power of transformers with chain of thought. In: The Twelfth International Conference on Learning Representations (2024)
18. Min, S., et al.: Rethinking the role of demonstrations: What makes in-context learning work? In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 11048–11064. Association for Computational Linguistics (2022)
19. OpenAI: Gpt-4 technical report (2024), <https://arxiv.org/abs/2303.08774>
20. Pan, J., et al.: What in-context learning “learns” in-context: Disentangling task recognition and task learning. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 8298–8319. Association for Computational Linguistics (2023)
21. Saparov, A., He, H.: Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In: The Eleventh International Conference on Learning Representations (2023)
22. Schaeffer, R., Pistunova, K., et al.: Invalid logic, equivalent gains: The bizarreness of reasoning in language model prompting. arXiv preprint arXiv:2307.10573 (2023)
23. Si, C., Friedman, D., et al.: Measuring inductive biases of in-context learning with underspecified demonstrations. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11289–11310. Association for Computational Linguistics (2023)
24. Talmy, L.: Toward a cognitive semantics: Concept structuring systems, vol. 1. MIT press (2000)
25. Team, G.: Gemma 2: Improving open language models at a practical size (2024), <https://arxiv.org/abs/2408.00118>

26. Team, L.: The llama 3 herd of models (2024), <https://arxiv.org/abs/2407.21783>
27. Team, Q.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)
28. Wang, B., Min, S., et al.: Towards understanding chain-of-thought prompting: An empirical study of what matters. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2717–2739. Association for Computational Linguistics (2023)
29. Wang, X., Zhou, D.: Chain-of-thought reasoning without prompting. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
30. Wei, J., Wang, X., Schuurmans, et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 24824–24837. Curran Associates, Inc. (2022)
31. Wei, J., Wei, J., et al.: Larger language models do in-context learning differently. arXiv preprint arXiv:2303.03846 (2023)
32. Wu, Y., Wang, Y., Du, T., Jegelka, S., Wang, Y.: When more is less: Understanding chain-of-thought length in llms (2025), <https://arxiv.org/abs/2502.07266>
33. Yang, W., Ma, S., Lin, Y., Wei, F.: Towards thinking-optimal scaling of test-time compute for llm reasoning (2025), <https://arxiv.org/abs/2502.18080>
34. Ye, X., Iyer, S., et al.: Complementary explanations for effective in-context learning. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 4469–4484. Association for Computational Linguistics (2023)
35. Yoo, K.M., Kim, J., et al.: Ground-truth labels matter: A deeper look into input-label demonstrations. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 2422–2437. Association for Computational Linguistics (2022)
36. Zhou, D., Schärli, N., et al.: Least-to-most prompting enables complex reasoning in large language models. In: The Eleventh International Conference on Learning Representations (2023)