Break the Chain: Large Language Models Can be Shortcut Reasoners

Anonymous ACL submission

Abstract

Recent advancements in Chain-of-Thought (CoT) reasoning utilize complex modules but are hampered by high token consumption, limited applicability, and challenges in reproducibility. This paper conducts a critical evaluation of CoT prompting, extending beyond arithmetic to include complex logical and commonsense reasoning tasks, areas where standard CoT methods fall short. We propose the integration of human-like heuristics and shortcuts into language models (LMs) through "break the chain" strategies. These strategies disrupt traditional CoT processes using controlled variables to assess their efficacy. Additionally, we develop innovative zero-shot prompting strategies that encourage the use of shortcuts, enabling LMs to quickly exploit reasoning clues and bypass detailed procedural steps. Our comprehensive experiments across various LMs, both commercial and open-source, reveal that LMs maintain effective performance with "break the chain" strategies. We also introduce ShortcutQA, a dataset specifically designed to evaluate reasoning through shortcuts, compiled from competitive tests optimized for heuristic reasoning tasks such as forward/backward reasoning and simplification. Our analysis confirms that ShortcutQA not only poses a robust challenge to LMs but also serves as an essential benchmark for enhancing reasoning efficiency in AI.

1 Introduction

In the evolving landscape of artificial intelligence, the ability to reason and solve complex problems symbolizes a cornerstone of intelligence. Language Models (LMs), particularly those based on transformer (Vaswani et al., 2017) architectures, have revolutionized our approach to natural language processing (NLP), significantly enhancing capabilities in comprehending and generating text that bears a striking resemblance to human communication.

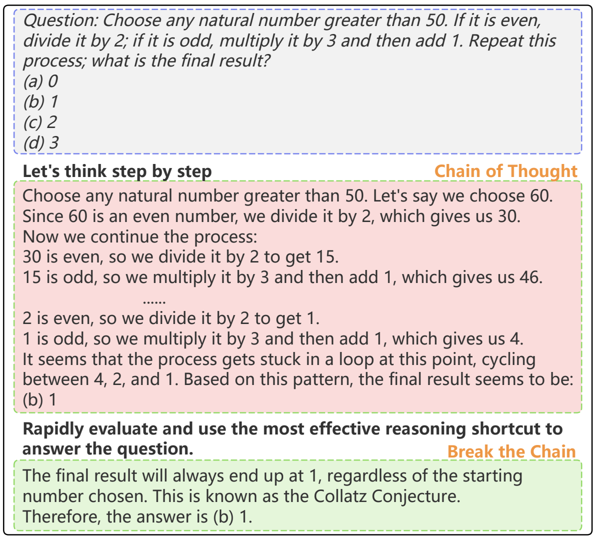

Figure 1: ChatGPT responses to Chain-of-Thought and "Break the Chain". Our "Break the Chain" method significantly simplifies the reasoning process.

Among recent advancements, Chain-of-Thought (CoT) prompting has emerged as a pivotal technique for utilizing Large Language Models (LLMs) to address complex reasoning tasks. By methodically eliciting step-by-step reasoning, CoT prompting has significantly enhanced the problem-solving capabilities of LLMs across a variety of learning scenarios, including few-shot (Wei et al., 2022) and zero-shot contexts (Kojima et al., 2022a). Figure 1 illustrates a zero-shot example in which the ChatGPT model methodically resolves a mathematical question. This strategy is further augmented by approaches such as self-consistency (Wang et al., 2022b, 2023c), interactive reasoning (Yao et al., 2022a; Shinn et al., 2024), reflective thinking (Ling et al., 2024; Li et al., 2023), task decomposition (Khot et al., 2022; Press et al., 2022), and strategic planning (Wang et al., 2023b; Hu et al., 2023).

Despite its benefits, CoT is also critiqued for its substantial token usage, as it explores numerous

[*]These authors contributed equally to this work
[†]Corresponding author

reasoning pathways before arriving at a conclusive answer. This characteristic is particularly prominent in variants such as Tree-of-Thought (ToT) (Yao et al., 2023), which scrutinize every possible reasoning chain. Traditionally, CoT has been predominantly applied to mathematical reasoning, with scant application to commonsense, or complex logical reasoning tasks. This limited focus may hinder a comprehensive understanding of CoT's potential to emulate intricate human-like reasoning processes. Additionally, instruction fine-tuned (IFT) (Ouyang et al., 2022) large language models like ChatGPT, which are usually capable of reaching the answers methodically, further question the necessity for explicit CoT prompting (Chen et al., 2023).

Human reasoning uses heuristics to find local rational maximum (Karlan, 2021; Neth and Gigerenzer, 2015; Lancia et al., 2023), which often relies on cognitive shortcuts (Fernbach and Rehder, 2013; Ferrario, 2004), a characteristic that can be mirrored and exploited in LMs. Traditionally, LLMs' shortcut learning has been viewed as the acquisition of spurious correlations within datasets (Du et al., 2023; Jiang and Bansal, 2019; Branco et al., 2021). However, this perspective fails to capture the nuanced heuristic reasoning processes inherent in human cognition, both in everyday scenarios and professional contexts such as clinical decision-making. We argue that shortcut reasoning, by drastically reducing reasoning steps and computational demands, offers a valuable means of enhancing LLM efficiency. As depicted in Figure 1, when prompted with shortcut reasoning, the ChatGPT model swiftly arrives at answers with minimal token consumption. The ability of LLMs to employ shortcut reasoning not only mirrors human cognitive strategies but also has the potential to streamline problem-solving processes, thereby reshaping computational efficiency and model performance.

The primary goal of our study is to critically evaluate and challenge the established Chain-of-Thought (CoT) prompting framework used in Large Language Models (LLMs). Our approach is three-pronged: First, we explore the effectiveness, limitations, and mechanisms of CoT by comparing it with different prompts derived from the "break the chain" strategy in both few-shot and zero-shot scenarios. Second, the study pioneers the use of shortcut reasoning prompts that encourage LLMs to utilize heuristic shortcuts — akin to intuitive leaps in human reasoning — to efficiently solve problems. This method aims to minimize computational demands and token consumption while maintaining or potentially enhancing performance accuracy. To support this investigation, we introduce ShortcutQA, a novel dataset meticulously curated to specifically assess the ability of LLMs to employ heuristic shortcuts. We conducted experiments on both OpenAI models and open-source models of various sizes, including MIXTRAL-8X7B-INSTRUCTION, LLAMA-3-70B-INSTRUCTION, QWEN1.5-72B-CHAT, QWEN1.5-14B-CHAT, QWEN1.5-1.8B-CHAT, to ascertain the generalizability of our experimental conclusions across different model configurations.

Our few-shot experiments reveal that Large Language Models (LLMs) are not adversely affected by disrupted Chain-of-Thought (CoT) demonstrations, casting doubts on the effectiveness of few-shot CoT methods. To our knowledge, this is the first series of experiments designed to "break the chain" of in-context examples. Furthermore, in zero-shot scenarios, models prompted with shortcut reasoning display robust performance, often surpassing that of traditional CoT methods. Our evaluations span both OpenAI models and open-source models, showing consistent results across platforms.

Furthermore, our comparative analysis elucidates distinct performance trends across various model sizes: smaller models typically experience more substantial enhancements with Chain-of-Thought (CoT) prompts compared to their larger counterparts. Notably, as model size increases, the efficacy of "break the chain" strategies becomes more pronounced, highlighting its effectiveness in mitigating the impact of disrupted CoT demonstrations.

Most notably, we observe that shortcut reasoning significantly reduces token consumption, providing a vital advantage in computational efficiency. Under stringent token constraints, shortcut reasoning strategies not only conserve resources but also consistently outperform traditional CoT methods. These benefits are observed across various datasets, underscoring the robustness and scalability of shortcut reasoning as a superior approach in enhancing LLM performance.

2 Related Work

2.1 CoT Prompting in Large Language Models

The evolution of Chain-of-Thought (CoT) prompting, particularly through few-shot (Wei et al., 2022) and zero-shot (Kojima et al., 2022a) methodologies, has markedly advanced Large Language Models' (LLMs) ability to address complex reasoning challenges. This field has witnessed the introduction of sophisticated data structures, such as Tree-of-Thought (Yao et al., 2023), Graph-of-Thought (Besta et al., 2024), and Program-of-Thought (Chen et al., 2022), enriching LLMs' capacity for introspection and nuanced evaluation of their reasoning paths.

Beyond conventional prompting strategies, the ReAct model (Yao et al., 2022b) integrates reasoning with actionable tasks like data retrieval, whereas the Selection-Inference framework (Creswell et al., 2023) combines context creation with logical chaining. While pioneering, these approaches rely on the models' inherent abilities and do not embed explicit logical rules within the reasoning process.

The adoption of external tools in prompting paradigms, especially for tasks that demand supplementary knowledge, has also shown considerable progress. Analogous to the role calculators play in mathematical reasoning, introducing predefined functions for enforcing inference rules marks a significant step forward in leveraging external computational aids to bolster reasoning capabilities.

Moreover, breaking down complex reasoning tasks into more manageable subproblems or engaging multiple models for collaborative problem-solving has introduced novel methodologies in LLM prompting. Strategies such as Cumulative Reasoning (Zhang et al., 2023a) focus on an iterative, step-wise approach, while ScratchPad (Nye et al., 2021) emphasizes the articulation of intermediate steps in multi-step reasoning. Meta-prompting (Suzgun and Kalai, 2024) envisions a cooperative framework where LLMs act as orchestrators, decomposing tasks, delegating them to specialized models, and synthesizing the outcomes, thereby fostering a holistic approach to problem-solving.

In the specific arena of instruct-tuning LLMs with tailored datasets for advanced reasoning, initiatives like LogiCoT (Liu et al., 2023), which fine-tunes an LLaMA-7b model with data on logical chaining, demonstrate considerable improvements in logical reasoning tasks. Similarly, LogicLLM (Jiao et al., 2023) explores a self-supervised learning strategy for logical reasoning enhancements, and Symbol-LLM (Xu et al., 2023) incorporates symbolic data in a two-stage tuning process to equip a LLaMA-2-chat model with symbolic reasoning skills. These efforts highlight the potential of fine-tuning with specialized datasets to significantly enhance the reasoning capabilities of LLMs, illustrating the dynamic and evolving landscape of CoT prompting in AI research.

2.2 Questioning CoT

Despite the demonstrated effectiveness of Chain-of-Thought (CoT) in enhancing model performance on complex tasks, the underlying mechanisms by which Large Language Models (LLMs) generate CoT responses are not fully understood. Research efforts are increasingly focused on demystifying CoT prompting, providing empirical insights and developing theoretical frameworks to comprehend this advanced reasoning capability. However, numerous studies have highlighted the brittleness of CoT reasoning in various aspects.

Turpin et al. (2023) investigate the faithfulness of CoT reasoning, revealing systematic misrepresentations in the true rationale behind a model's predictions. Lanham et al. (2023) extend this inquiry by introducing errors or paraphrases within the CoT process to test whether the articulated reasoning truly reflects the model's underlying logic, finding that larger models tend to produce more unfaithful responses. This issue of faithfulness is critical as it challenges the reliability of CoT explanations. The effectiveness of CoT is also impacted by the selection and arrangement of demonstrations. Wang et al. (2023a) find that the accuracy of reasoning chains is less critical than the relevance of the question and the correctness of the reasoning sequence, emphasizing the importance of contextual alignment. In contrast, Wang et al. (2022a) show that CoT can operate even with invalid demonstrations, suggesting some resilience in the reasoning process. Our research contributes to this discourse by disturbing the order of the reasoning chain to examine its impact on CoT consistency.

Jin et al. (2024) demonstrate that artificially lengthening the reasoning steps in prompts — simply by instructing models to "think more steps" — can enhance LLMs' performance across various datasets without introducing new content. This

Dataset	Question Type	# of instances	Avg. # words	Source
ShortcutQA	Analytical shortcuts	156	55.88	Analytical reasoning tests
	Logical shortcuts	108	21.76	Verbal reasoning tests
	Mathematical shortcuts	185	67.19	Gaokao examinations

Table 1: Dataset statistics of ShortcutQA.

finding suggests that the perceived depth of reasoning may artificially inflate effectiveness. Conversely, we explore minimalist prompting strategies where LLMs are instructed to streamline their reasoning processes.

The sensitivity of LLMs to the ordering of premises is scrutinized by Chen et al. (2024), who note optimal performance when the order of premises supports the necessary context in intermediate reasoning steps. This sensitivity is paradoxical in deductive reasoning contexts where the order of premises should not logically influence the validity of conclusions. Similarly, Pfau et al. (2024) Indicates that LLMs solve more problems with meaningless filler tokens in place of a chain of thought than without meaningless tokens. This finding suggests that CoT's effectiveness may sometimes rely solely on the increase in computational effort, rather than on the literal intermediate reasoning steps. Our "break the chain" methods experiment with new models and datasets and aim to illuminate this issue further.

Implicit CoT (Deng et al., 2023, 2024) has been introduced to internalize explicit step-by-step reasoning. Similar to our work, implicit CoT questions the necessity of step-by-step reasoning. However, we diverge from prior studies that employed fine-tuning to reduce the need for reasoning steps.

Finally, Chen et al. (2023) question the applicability of CoT in instruction fine-tuned (IFT) models like ChatGPT, which show inconsistent performance across various reasoning tasks. Surprisingly, while CoT prompts enhance some reasoning tasks, they fail in others like arithmetic reasoning, where ChatGPT can independently generate CoT sequences without specific prompts. This phenomenon inspires us to abstract a hypothesis that more powerful models increasingly exhibit a reduced dependency on CoT. Our subsequent experiments conducted within the Qwen1.5 series of various sizes strive to support this viewpoint.

3 ShortcutQA

The ShortcutQA dataset is designed to evaluate Language Models' (LMs) ability to employ heuristic shortcuts in reasoning, addressing a gap in existing resources that primarily focus on sequential reasoning approaches. Comprising 449 diverse reasoning problems, ShortcutQA spans logical puzzles to real-world problem-solving scenarios. Each problem is presented with a shortcut-based solution alongside a detailed step-by-step solution, categorized into three reasoning types.

Data Collection and Annotation

Data for ShortcutQA were sourced from various online forums and educational websites, with necessary permissions secured. Annotation was conducted by two independent domain experts, adhering to strict guidelines for identifying and categorizing heuristic shortcuts employed in the solutions. A third expert resolved any discrepancies, ensuring high annotation quality and consistency.

Dataset Categorization

ShortcutQA introduces problems categorized into three distinct types, each testing different aspects of heuristic reasoning:

- **Analytical Shortcuts:** Tasks necessitate analyzing situations beyond mere comprehension, assessing models' capabilities in efficiently synthesizing and utilizing key information, and strategic decision-making under time constraints.

- **Logical Shortcuts:** Encompassing forms of reasoning such as analogical, abductive, and forward/backward reasoning, these tasks focus on applying these logical theories to derive conclusions from provided statements.

- **Mathematical Shortcuts:** Features problems solvable through approximation techniques, substitution, simplification, and special-case reasoning, bypassing traditional sequential thought processes.

Data Statistics are shown in Table 1. We release the data at https://anonymous.com.

4 Method

4.1 Break the Chain

To examine the resilience and limitations of Large Language Models (LLMs) in employing Chain-of-Thought (CoT) reasoning, our research outlines a novel experimental framework aimed at "breaking the chain" of thought. This approach seeks to elucidate the conditions under which CoT reasoning may falter, thereby offering insights into the underlying mechanisms of LLMs' reasoning capabilities. Our methodology juxtaposes zero-shot and few-shot scenarios to delineate the impact of CoT disruption across different prompting contexts.

Few-Shot In the few-shot scenario, our strategy involves perturbing the sequence of sentences within the in-context examples provided to the LLM. This disturbance is designed to misalign the logical progression typically demonstrated in CoT reasoning, thereby testing the model's ability to maintain coherent and accurate reasoning despite the disordered presentation of steps. This manipulation will help ascertain the significance of stepwise logical progression in the model's reasoning efficacy and its ability to reorient itself to reach correct conclusions.

Zero-Shot We initiate probing experiments to assess the efficacy of zero-shot CoT prompts, aiming to discern whether CoT prompting is essential or merely a byproduct of longer model responses. Employing controlled experiments, we craft prompts that obviate the need for reasoning chains, instructing models to provide either more verbose or minimalist responses. Detailed descriptions of these prompts are provided in Appendix A. Furthermore, we employ meticulously designed prompts to stimulate shortcut reasoning, outlined comprehensively in Appendix A. By directing LLMs to circumvent intermediate reasoning steps typically associated with CoT, we aim to evaluate the resilience of their inferential processes and their reliance on detailed reasoning pathways.

ShortcutQA Probing Parallel to our few-shot and zero-shot experiments, we introduce the ShortcutQA dataset into our methodology. ShortcutQA is carefully curated to focus on questions that require shortcut reasoning — a form of intuitive problem-solving that deviates from traditional step-by-step logical deduction. The inclusion of ShortcutQA is intended to test the hypothesis that LLMs can effectively employ heuristic shortcuts, akin to human cognitive shortcuts, to efficiently resolve complex problems.

4.2 Experimental Setup

We evaluate Large Language Models (LLMs) across a variety of commercial and open-source platforms under both few-shot and zero-shot conditions. Our methodology includes a diverse array of complex problem-solving tasks encompassing arithmetic reasoning, commonsense deduction, and logical reasoning. This design rigorously tests the LLMs' ability to generalize across different difficulty levels and domains.

Task	Dataset	Size	Avg #words
Arithmetic	SingleEq	508	27.4
	AddSub	395	31.5
	MultiArith	600	31.8
	GSM8K	1319	46.9
	AQUA-RAT	254	51.9
	SVAMP	1000	31.8
Commonsense	CommonsenseQA	1221	27.8
	StrategyQA	2290	9.6
Logic	Date Understanding	369	35.0
	Coin Flip	500	37.0
	LogiQA	651	146.2
	ReClor	500	153.0

Table 2: Statistics of Evaluation benchmarks.

As depicted in Figure 5 in Appendix B, the experimental pipeline begins by inputting a question and a prompt into an LLM, which then generates a reasoned response and answer. This output is concatenated with the original question and prompt, followed by an answer extraction prompt to extract the final answer.

Benchmarks For *arithmetic reasoning*, we assess the models using six datasets: SingleEq (Koncel-Kedziorski et al., 2015), AddSub (Hosseini et al., 2014) , MultiArith (Roy and Roth, 2015) , GSM8K (Cobbe et al., 2021), AQUA-RAT (Ling et al., 2017), and SVAMP (Patel et al., 2021). The first three originate from the well-established Math World Problem Repository (Koncel-Kedziorski et al., 2016), with the remaining datasets presenting more recent and complex challenges. SingleEq and AddSub feature relatively straightforward problems that can be solved without multi-step reasoning, whereas MultiArith, AQUA-RAT, GSM8K, and SVAMP require more intricate, sequential problem-solving.

For *commonsense reasoning*, we utilize the CommonsenseQA (Talmor et al., 2019) and StrategyQA (Geva et al., 2021) datasets. CommonsenseQA tests reasoning based on general world knowledge (Talmor et al., 2019) , while StrategyQA demands inference of unstated, multi-step reasoning processes (Geva et al., 2021).

For *logical reasoning tasks*, we select two scenarios from the BIG-bench (Srivastava et al., 2022): Date Understanding and Coin Flip (Wei et al.,

Task	Dataset	Few-shot		Zero-shot		
		Base	Break the Chain	Base	No Steps	More Tokens
Arithmetic	SingleEq	**92.72**	92.32	86.61	**90.35**	88.39
	AddSub	84.05	**85.32**	83.80	**89.62**	86.58
	MultiArith	**99.00**	98.33	83.33	91.17	**93.50**
	GSM8K	**74.60**	74.22	32.68	37.53	**38.89**
	AQUA-RAT	53.15	**55.51**	35.43	36.61	**38.97**
	SVAMP	76.80	**79.70**	71.70	**81.70**	76.70
Commonsense	CommonsenseQA	74.94	**75.18**	70.52	**75.92**	74.28
	StrategyQA	**69.13**	68.60	64.37	59.91	**63.23**
Logic	Date Understanding	81.03	**82.11**	64.37	**64.50**	63.23
	LogiQA	**35.94**	33.95	40.09	**41.17**	40.09
	ReClor	**51.40**	50.80	52.40	51.20	**54.20**

Table 3: ChatGPT performance comparison across tasks. All results are in %, the best ones are in **bold**.

2022). Date Understanding challenges models to infer dates from given contexts, and Coin Flip evaluates the ability to determine the outcome of a series of coin flips. Additionally, we incorporate LogiQA (Liu et al., 2020) and ReClor (Yu et al., 2020), which are reading comprehension tests that require logical deduction.

Language Models We test both OpenAI commercial models and huggingface open-source models. For OpenAI models, we choose the ChatGPT (gpt-3.5-turbo-0613) model, an IFT GPT-3 model. For community models, we use Llama-3-70B-Instruct, Mixtral-8x7B-Instruct, Qwen1.5-72B-Chat, Qwen1.5-14B-Chat, Qwen1.5-1.8B-Chat.

Baselines We run zero-shot CoT (Kojima et al., 2022b) and few-shot CoT (Wei et al., 2022) on the datasets to establish our baselines. In the few-shot CoT setup, we follow Wei et al. (2022) to provide each test with context examples; for the zero-shot baseline, each question is suffixed with "The answer is ", following prior work (Kojima et al., 2022b; Zhang et al., 2023b).

5 Results

Few-Shot Table 3 on the left side shows comparative performance between traditional few-shot CoT and our "breaking the chain" approach across datasets in commonsense, arithmetic, and logical reasoning tasks. Notably, in arithmetic reasoning, performance on the MultiArith dataset decreases slightly from 99.00% to 98.33% with "breaking the chain", while in GSM8K, the decrease is marginal, from 74.60% to 74.22%. In commonsense reasoning, "breaking the chain" slightly outperforms the traditional approach on CommonsenseQA (75.18% vs. 74.94%), but underperforms on StrategyQA, dropping from 69.13% to 68.60%. LogiQA in logical reasoning shows a more notable performance drop from 35.94% to 33.95%. These results suggest that while "breaking the chain" generally performs comparably to the few-shot CoT baseline, it does not significantly impact the model's overall performance.

Zero-Shot The right side of Table 3 presents results from our zero-shot probing experiment, comparing the zero-shot CoT baseline with our "break the chain" prompts across 11 datasets within three key tasks: arithmetic reasoning, commonsense reasoning, and logical reasoning. Notably, even when we ablate step-by-step reasoning, ChatGPT maintains competitive performance across various tasks. Moreover, prompting with only "More Tokens" leads to the best performance on several other datasets.

Results for the "Shortcut Reasoning" prompts are detailed in Table 4, where this approach shows substantial improvements: a 22% increase in arithmetic tasks, a 9% boost in commonsense tasks, and an 11% enhancement in logical reasoning tasks. Performance is consistent on the Mixtral and Qwen platforms, though it varies with the Llama models, underlining the effectiveness of our approach.

In addition, experiments with Qwen models of varying sizes, both under CoT and "break the chain" conditions, are documented. Figure 4 in Appendix C illustrates that smaller models exhibit a more pronounced reliance on CoT, especially as the model size decreases, narrowing performance gaps from a 16% deficit in 72B models to parity in 1.8B models for arithmetic tasks. For logic and commonsense tasks, smaller models transition from underperformance to outperforming larger counterparts, suggesting less capable models benefit more from CoT's structured approach.

Model	Task	Base	Quick Conclude	Shortcut Reasoning	Effective Shortcut	Innovative Shortcut
ChatGPT	Arithmetic	65.59	77.23	80.11	**80.58**	71.34
	Commensense	67.45	73.18	**73.65**	72.36	67.52
	Logical	51.97	53.32	**57.57**	56.77	56.91
Llama-70B	Arithmetic	72.47	62.29	**81.59**	63.37	50.96
	Commensense	67.57	**73.00**	60.58	67.29	67.27
	Logical	**71.41**	66.26	68.60	67.18	63.95
Mixtral-8x7B	Arithmetic	70.80	**73.22**	71.70	68.77	56.63
	Commensense	65.03	69.37	69.23	**69.50**	60.81
	Logical	69.08	69.61	**69.84**	68.48	60.46
Qwen1.5-72B	Arithmetic	65.28	**76.00**	75.51	74.52	70.83
	Commensense	79.11	79.85	79.38	**80.36**	79.78
	Logical	60.17	63.42	63.58	**63.62**	61.79
Qwen1.5-14B	Arithmetic	63.30	**71.97**	71.57	69.94	66.85
	Commensense	74.43	**75.65**	75.14	75.25	74.20
	Logical	53.47	**55.76**	54.91	55.50	54.38
Qwen1.5-1.8B	Arithmetic	39.40	**39.99**	37.12	31.97	28.81
	Commonsense	**57.61**	55.07	57.19	55.95	55.20
	Logical	**33.00**	30.72	31.00	32.37	32.05

Table 4: Experiment results concerning different tasks. Detailed results are in Appendix C. All results are in %, the best ones are in **bold**.

These findings question the prevailing assumption that CoT invariably enhances LLM performance. Our results indicate that specific prompts, even without detailed reasoning, can yield comparable or superior outcomes. However, the effectiveness of "break the chain" prompts varies, pointing to a nuanced interplay between prompt nature and LLM performance that merits further investigation.

We observe that CoT is particularly adept at tackling questions decomposable into sub-issues that are solvable in brief sentences. Challenges arise when generated responses become excessively lengthy, leading to potential task misalignment and illogical outputs, or when they exceed the maximum length constraints set in the code, inhibiting the completion of reasoning sequences.

ShortcutQA Table 5 presents a comparative analysis of performance across various task types within the ShortcutQA dataset. Compared to benchmarks utilized elsewhere in this study, ShortcutQA poses a greater challenge, making it an ideal testing ground for advancing model capabilities.

In mathematical reasoning tasks, all "break the chain" prompts outperform the established baselines. The "Innovative Shortcut" prompt is particularly effective, achieving a significant relative improvement of 28.56% over the baseline. "Quick Conclude" also shows substantial gains, with a relative increase of 23.8% compared to the baseline.

For analytical and verbal reasoning tasks, "Quick Conclude" registers the highest improvements, with increases of 26.65% and 9.99%, respectively, over the baseline. "Innovative Shortcut" also posts notable gains in analytical tasks, while "Effective Shortcut" sees considerable enhancements in verbal tasks.

Overall, "Innovative Shortcut" and "Quick Conclude" are standout performers on the ShortcutQA dataset, underscoring the potency of our "break the chain" strategy. This dataset not only challenges current LLMs but also sets a benchmark for future enhancements, providing a robust platform for testing and refining next-generation models.

6 Discussion

6.1 Reasoning with Token Limits

We investigated the impact of token limits on model performance by experimenting with different constraints (128, 256 tokens) during the response generation phase. Figure 2 illustrates how varying token limits affect outcomes on the mathematical reasoning task within ShortcutQA using different prompts. We observed that as the token limit increases, so does performance across all prompts, indicating that constraints on output length significantly influence the inference process and thus the results. Notably, even at the minimum limit of 128 tokens, all prompts exceed the baseline performance, suggesting that our "break the chain" approach is not only efficient but also effective in

Dataset	Question Type	Base	Quick Conclude	Shortcut Reasoning	Effective Shortcut	Innovative Shortcut
ShortcutQA	Analytical Reasoning	26.79	**33.93**	21.43	19.64	30.36
	Verbal Reasoning	22.73	**25.00**	22.73	23.86	21.59
	Mathematical Reasoning	25.00	30.95	29.76	26.19	**32.14**

Table 5: Performance comparison across tasks within ShortcutQA.

conserving computational resources while maintaining or improving task performance.

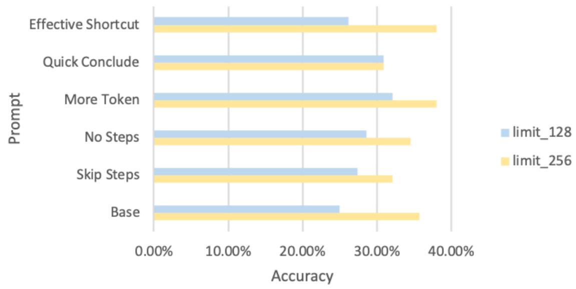

Figure 2: Performance comparison of different token limits on the mathematical reasoning task from ShortcutQA.

6.2 Theoretical Analysis

We have developed a qualitative model to formalize the performance dynamics of Chain-of-Thought (CoT) reasoning and to elucidate the effectiveness of the "Break the Chain" approach.

In our framework, each CoT step is divided into two phases: analysis and reasoning. The accuracy of the analysis at step t is denoted as $P(a_t)$, and the subsequent reasoning based on this analysis is denoted as $P(r_t)$. Therefore, the total accuracy of a CoT sequence depends on the combined accuracy of these phases across all steps, mathematically expressed as:

$$P(\text{CorrectReasoning}_{\text{CoT}}) = \prod_{t=1}^{T} P(a_t)P(r_t), \quad (1)$$

where T is the total number of steps in the CoT reasoning chain. To evaluate the efficacy of different prompting strategies, we define $P(\text{CorrectReasoning}_p)$ as the probability of achieving correct reasoning for a given prompt p. A prompt is considered more effective than the traditional CoT approach if $P(\text{CorrectReasoning}_p)$ surpasses $P(\text{CorrectReasoning}_{\text{CoT}})$.

In cases where no explicit analysis or reasoning phase is involved, and both are integrated by LLMs in each step, Equation 1 simplifies to:

$$P(\text{CorrectReasoning}_{\text{CoT}}) = \prod_{t=1}^{T} P(i_t), \quad (2)$$

where i_t signifies the probability of obtaining the correct result in a single, consolidated inference step.

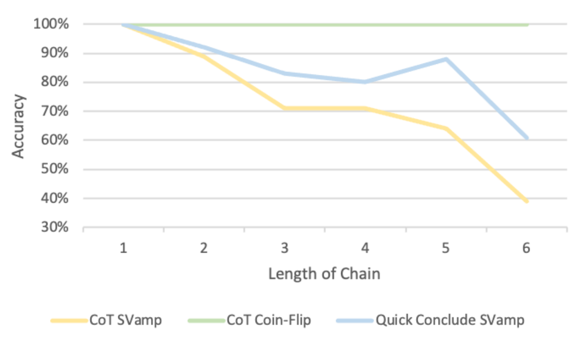

Figure 3: Relationship between CoT Chain Length and Accuracy.

Our experimental results corroborate the theoretical predictions, as illustrated in Figure 3. We observe that CoT accuracy generally declines as chain length increases. Notably, in scenarios like Coin Flip where $P(i_t)$ approaches 1, accuracy remains stable regardless of chain length. Conversely, in tasks like SVamp where $P(i_t)$ is lower, a decrease in accuracy is noted as the chain lengthens. When comparing "Quick Conclude" on SVamp against baseline accuracies, the relative CoT accuracy diminishes with increasing chain length, aligning precisely with our model. Detailed methodologies for these experiments are available in Appendix D.

7 Conclusion

This study critically evaluates Chain-of-Thought (CoT) reasoning in language models, highlighting limitations such as high token consumption and limited applicability. Our "break the chain" strategies integrate human-like heuristics and shortcuts, enhancing efficiency without compromising performance across various models. The introduction of the ShortcutQA dataset further advances AI reasoning evaluation by focusing on heuristic tasks, providing a robust benchmark that challenges traditional methods. Our findings suggest that adopting more intuitive, efficient reasoning approaches could significantly improve the problem-solving capabilities of AI systems in real-world applications.

Limitations

While our study presents a significant advancement in understanding the reasoning capabilities of Large Language Models (LLMs) through the introduction of "break the chain" strategies and the ShortcutQA dataset, there are several limitations that warrant discussion.

1. Scope of Reasoning Tasks: Our experiments, although diverse, are not exhaustive in terms of the types of reasoning tasks. The tasks selected for our study are primarily logical, mathematical, and commonsense reasoning problems. There may be other types of reasoning tasks where the "break the chain" approach could exhibit different performance characteristics.

2. Faithfulness of Reasoning: As noted in related work, there is an ongoing debate regarding the faithfulness of CoT reasoning in LLMs. Our study raises questions about the necessity of explicit step-by-step reasoning, but does not fully resolve the issue of whether LLMs can provide explanations that are both accurate and reflective of their internal reasoning processes.

3. Evaluation Metrics: Our evaluation primarily relies on accuracy as the metric for assessing reasoning performance. However, reasoning effectiveness may also be influenced by other factors such as the coherence, explainability, and efficiency of the reasoning process, which were not extensively measured in this study.

In future work, it will be crucial to address these limitations by expanding the scope of reasoning tasks, investigating the generalizability of the strategies across different model architectures, mitigating potential biases in the dataset, exploring different token constraints, enhancing the faithfulness of reasoning, and considering a broader set of evaluation metrics. Furthermore, research into the practical application of these strategies in real-world scenarios will be essential to fully harness the potential of LLMs as efficient and effective reasoners.

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michał Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*.

Ruben Branco, António Branco, Joao Rodrigues, and Joao Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521.

Jiuhai Chen, Lichang Chen, Heng Huang, and Tianyi Zhou. 2023. When do you need chain-of-thought prompting for chatgpt? *arXiv preprint arXiv:2304.03262*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Xinyun Chen, Ryan A Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. *arXiv preprint arXiv:2402.08939*.

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. *CoRR*, abs/2110.14168.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*.

Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024. From explicit cot to implicit cot: Learning to internalize cot step by step. *Preprint*, arXiv:2405.14838.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. 2023. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *Commun. ACM*, 67(1):110–120.

Philip M Fernbach and Bob Rehder. 2013. Cognitive shortcuts in causal inference. *Argument & Computation*, 4(1):64–88.

Catherine G Ferrario. 2004. Developing clinical reasoning strategies: cognitive shortcuts. *Journal for Nurses in Professional Development*, 20(5):229–235.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Preprint*, arXiv:2101.02235.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization.

In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 523–533. ACL.

Hanxu Hu, Hongyuan Lu, Huajian Zhang, Wai Lam, and Yue Zhang. 2023. Chain-of-symbol prompting elicits planning in large langauge models. *arXiv preprint arXiv:2305.10276*.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop qa. *arXiv preprint arXiv:1906.07132*.

Fangkai Jiao, Zhiyang Teng, Shafiq Joty, Bosheng Ding, Aixin Sun, Zhengyuan Liu, and Nancy F. Chen. 2023. Logicllm: Exploring self-supervised logic-enhanced training for large language models. *Preprint*, arXiv:2305.13718.

Mingyu Jin, Qinkai Yu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, Mengnan Du, et al. 2024. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

Brett Karlan. 2021. Reasoning with heuristics. *Ratio*, 34(2):100–108.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022a. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022b. Large language models are zero-shot reasoners. In *NeurIPS*.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California. Association for Computational Linguistics.

Gian Luca Lancia, Mattia Eluchans, Marco D'Alessandro, Hugo J Spiers, and Giovanni Pezzulo. 2023. Humans account for cognitive costs when finding shortcuts: An information-theoretic analysis of navigation. *PLOS Computational Biology*, 19(1):e1010829.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *Preprint*, arXiv:2307.13702.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Heng Huang, Jiuxiang Gu, and Tianyi Zhou. 2023. Reflection-tuning: Data recycling improves llm instruction-tuning. *arXiv preprint arXiv:2310.11716*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 158–167. Association for Computational Linguistics.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2024. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 36.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023. Logicot: Logical chain-of-thought instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2908–2921.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3622–3628. ijcai.org.

Hansjörg Neth and Gerd Gigerenzer. 2015. Heuristics: Tools for an uncertain world. In *Emerging trends in the social and behavioral sciences*, pages 1–18. Wiley Online Library.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

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

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *Preprint*, arXiv:2103.07191.

Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. Let's think dot by dot: Hidden computation in transformer language models. *Preprint*, arXiv:2404.15758.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv preprint arXiv:2401.12954*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

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

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2022a. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2717–2739. Association for Computational Linguistics.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023c. Scott: Self-consistent chain-of-thought distillation. *arXiv preprint arXiv:2305.01879*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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

Fangzhi Xu, Zhiyong Wu, Qiushi Sun, Siyu Ren, Fei Yuan, Shuai Yuan, Qika Lin, Yu Qiao, and Jun Liu. 2023. Symbol-llm: Towards foundational symbol-centric interface for large language models. *Preprint*, arXiv:2311.09278.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022a. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. *ArXiv*, abs/2210.03629.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2023a. Cumulative reasoning with large language models. *ArXiv*, abs/2308.04371.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

A Zero-shot prompts for "break the chain"

The abbreviations of probing prompts and shortcut prompts are shown in the table 6.

B Pipeline Details

Figure 5 shows the pipeline of experiments.

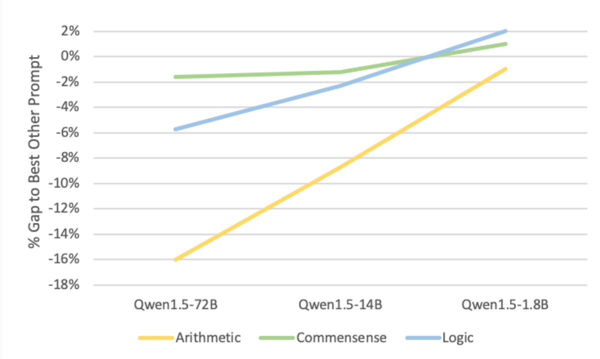

Figure 4: The Impact of Model Size on CoT's Relative Outperformance over Other Prompts across Datasets

C Experiment Results

Table 7 is the original experiment results of diverse model structures.

Figure 4 shows that as model size decreases, CoT's relative performance advantage over other prompts increases across all tasks.

D Detailed Methods

In this section, we introduce our detailed methods for our experiments. For our experiment in discussion, we generally used GPT4 to evaluate the logs, and caculate the accuracy of different lengths. First, We used GPT4 to check the logs of CoT to calculate the length of chain in each question on SVamp and Coin Flip. Second, we calculated the accuracy at different length of chain. Third, to exclude the disturbance of various difficulty distributions within each group, we calculated the accuracy with promptQC in each group of data as baseline without CoT on SVamp.

Prompt Type	Abbreviations	Full Prompts
Probing Prompts	Skip Steps	Let's skip as much as possible.
	No Steps	Let's don't think step by step.
	More Token	Let's think as much as possible.
Shortcut Prompts	Quick Conclude	Let's quickly conclude the answer without showing step-by-step reasoning.
	Shortcut Reasoning	Let's quickly conclude the answer with shortcut reasoning.
	Effective Shortcut	Rapidly evaluate and use the most effective reasoning shortcut to answer the question.
	Innovative Shortcut	Think outside the box and quickly identify an innovative shortcut to solve this problem.

Table 6: The relationship between a prompt abbreviation and its full prompt.

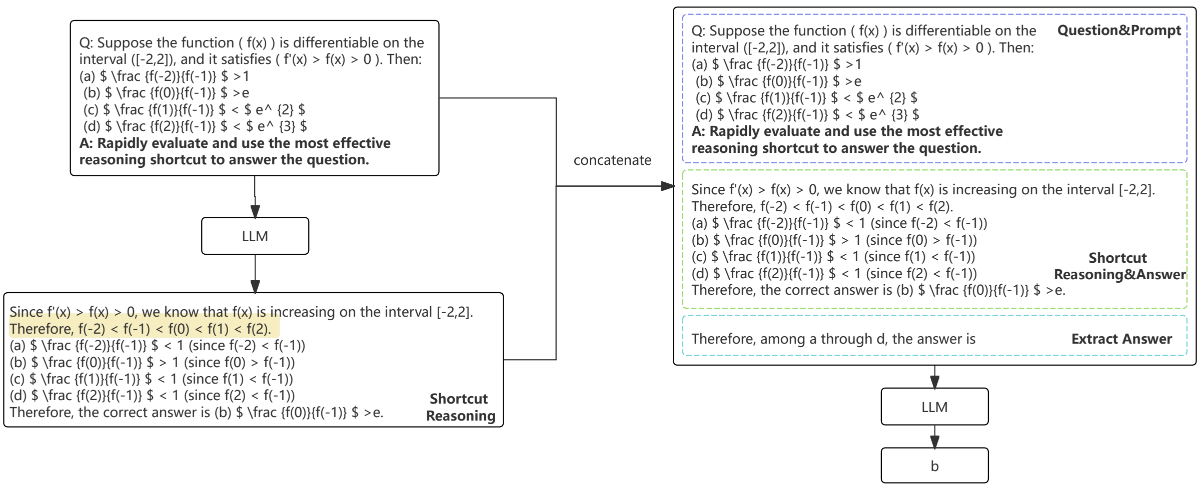

Figure 5: Our evaluation pipeline.

model	Task	Dataset	Base	Quick Conclude	Shortcut Reasoning	Effective Shortcut	Innovative Shortcut
ChatGPT	Arithmetic	SingleEq	86.61	91.14	91.73	**92.32**	77.36
		AddSub	83.80	**90.89**	86.33	89.62	73.67
		AQUA-RAT	35.43	51.97	52.76	**53.94**	52.36
		MultiArith	83.33	91.00	94.67	**94.83**	89.83
		GSM8K	32.68	56.86	**71.57**	67.25	58.00
		SVAMP	71.70	81.50	83.60	**85.50**	76.80
	Commensense	CommonsenseQA	70.52	77.89	**78.95**	76.82	72.89
		StrategyQA	64.37	**68.47**	68.34	67.90	62.14
	Logic	LogiQA	40.09	41.32	**43.32**	42.70	41.01
		ReClor	52.40	**52.80**	51.60	52.00	51.40
		Date Understanding	63.41	65.85	77.78	75.61	**78.32**
Llama-70B	Arithmetic	SingleEq	67.91	55.91	**81.10**	49.80	40.16
		AddSub	69.87	40.51	**80.25**	53.92	32.66
		AQUA-RAT	61.02	52.36	**62.20**	57.87	50.79
		MultiArith	79.83	71.00	**93.33**	73.67	60.33
		GSM8K	80.97	81.05	**85.14**	78.17	69.52
		SVAMP	75.20	72.90	**87.50**	66.80	52.30
	Commensense	CommonsenseQA	79.03	**81.49**	77.31	69.21	74.37
		StrategyQA	56.11	64.50	43.84	**65.37**	60.17
	Logic	LogiQA	57.60	57.30	57.45	**58.83**	56.07
		ReClor	**71.80**	69.40	68.40	69.80	70.20
		Date Understanding	**84.82**	72.09	79.95	72.90	65.58
Mixtral-8x7B	Arithmetic	SingleEq	87.40	**88.58**	87.01	83.46	70.87
		AddSub	85.82	**86.84**	84.81	83.54	72.15
		AQUA-RAT	37.40	41.34	**42.13**	39.76	32.68
		MultiArith	**87.50**	87.33	85.83	78.17	62.17
		GSM8K	48.90	**55.80**	54.81	49.20	39.12
		SVAMP	77.80	**79.40**	75.60	78.50	62.80
	Commensense	CommonsenseQA	71.63	72.40	**72.73**	71.17	65.85
		StrategyQA	58.43	66.33	65.72	**67.82**	55.76
	Logic	LogiQA	42.70	45.01	38.56	40.86	**45.78**
		ReClor	47.40	50.60	**51.80**	48.60	47.60
		Date Understanding	66.40	67.48	63.96	**68.02**	59.08
Qwen1.5-72B	Arithmetic	SingleEq	80.71	87.80	**88.78**	88.58	86.61
		AddSub	84.56	84.81	86.33	**88.61**	86.84
		AQUA-RAT	37.80	47.24	**48.43**	46.46	35.43
		MultiArith	81.33	**96.00**	95.33	**96.00**	93.67
		GSM8K	28.96	**54.06**	48.98	45.26	42.30
		SVAMP	78.30	**86.10**	85.20	82.20	80.10
	Commensense	CommonsenseQA	81.98	83.54	81.98	83.37	**83.7**
		StrategyQA	76.24	76.16	76.77	**77.34**	75.85
	Logic	LogiQA	46.54	50.08	**51.15**	50.54	47.00
		ReClor	61.60	**66.20**	64.00	65.80	64.40
		Date Understanding	72.36	73.98	**75.61**	74.53	73.98

Table 7: Comparison of Various Open-Source Large Models' Performance with Different Prompts Across Multiple Datasets.