Towards Better Chain-of-Thought: A Reflection on Effectiveness and Faithfulness

Anonymous ACL submission

Abstract

Chain-of-thought (CoT) prompting demonstrates varying performance under different reasoning tasks. Previous work attempts to evaluate it but falls short in providing an in-depth analysis of patterns that influence the CoT. In this paper, we study the CoT performance from the perspective of effectiveness and faithfulness. For the former, we identify key factors that influence CoT effectiveness on performance improvement, including problem difficulty, information gain, and information flow. For the latter, we interpret the unfaithful CoT issue by conducting a joint analysis of the information interaction among the question, CoT, and answer. The result demonstrates that, when the LLM predicts answers, it can recall correct information missing in the CoT from the question, leading to the problem. Finally, we propose a novel algorithm to mitigate this issue, in which we recall extra information from the question to enhance the CoT generation and evaluate CoTs based on their information gain. Extensive experiments demonstrate that our approach enhances both the faithfulness and effectiveness of CoT.

1 Introduction

Recently, with chain-of-thought (CoT) techniques (Wei et al., 2022), large language models (LLMs) are able to reason on complex tasks (Wang et al., 2023; OpenAI, 2023). By scaling the CoT process using reinforcement learning (RL), LLMs can even surpass human performance in competition-level mathematical problems (OpenAI, 2024; DeepSeek-AI et al., 2025). However, despite the significant success of the CoT, some studies find that it demonstrates poor performance on certain tasks (Sprague et al., 2024; Xu and Ma, 2024; Turpin et al., 2023; Lanham et al., 2023). In some cases, using CoT for the model's reasoning is unnecessary or even harmful (Wang et al., 2024b; Li et al., 2024).

These conflicting findings motivate the need for a systematic analysis of the CoT. To this end, a series of studies evaluating CoT's performance has commenced (Turpin et al., 2023; Bao et al., 2024; Wang et al., 2024b; Lanham et al., 2023), which can be mainly divided into two lines: On the one hand, some works assess CoT based on its effectiveness. They measure the accuracy improvements brought by the CoT across different tasks and identify task types where CoT is effective (Sprague et al., 2024; Xu and Ma, 2024; Madaan et al., 2023a). On the other hand, some works evaluate the CoT based on its faithfulness (Jacovi and Goldberg, 2020; Atanasova et al., 2023). They investigate the consistency between CoTs and final answers by analyzing the causal relevance linking them. (Lanham et al., 2023; Parcalabescu and Frank, 2023; Bao et al., 2024). Effectiveness is result-oriented, focusing on whether CoT can enhance the quality of reasoning outcomes; whereas faithfulness is process-oriented, concerned with whether the reasoning process of CoT genuinely influences the results.

Though these works have made great progress, they lack an in-depth analysis of the patterns influencing CoT performance. For the effectiveness evaluation works, they draw conclusions like CoT performs well in tasks involving mathematical symbols, but does not explore the underlying factors influencing these conclusions (Sprague et al., 2024; Xu and Ma, 2024). For the faithfulness evaluation works, they primarily design various methods to determine whether CoT is faithful, but lack an explanation for the issue of CoT unfaithfulness. (Lyu et al., 2023; Lanham et al., 2023; Bao et al., 2024).

In this paper, we focus on analyzing key patterns that influence the CoT's performance from both effectiveness and faithfulness perspectives. From the effectiveness perspective, we identify three factors that contribute to CoT's final improvement, including problem difficulty, information gain, and information flow. We start by splitting

questions into various difficulty levels and comparing the model's accuracy on them, from which we find that CoT is more effective on harder problems. Then, we calculate the information gain brought by CoT for questions across different tasks and demonstrate CoT with more additional information tends to be more effective. Lastly, we consider the internal information flow during model reasoning. Through the experiment, we conclude that the more information interaction increases with the CoT process, the more effective the CoT becomes. From the faithfulness perspective, we discover that there exist non-negligible unfaithful CoT issues in logical reasoning, where an incorrect CoT can still lead to the correct answer. We further interpret this issue by jointly analyzing the information interaction among question, CoT, and answer. Through it, we identify three patterns that lead to the CoT's unfaithfulness: (1) CoT loses key information from the question; (2) CoT transfers less information to the answer; (3) The model recalls correct information from the question when answering.

At last, we explore the relationship between the above two perspectives. A novel algorithm called **QU**estion **I**nformation **R**ecall and **E**nhancement (QUIRE) is designed to mitigate the unfaithful CoT issue. In it, we first generate a raw answer to recall correct information from the question, then use this extra information to prompt the generation of a new CoT generation. Finally, we employ the CoT information gain as the weight to vote for the final answer. Through extensive experiments, we not only demonstrate that our method can mitigate unfaithful issues, but also show that CoT faithfulness is a key factor in influencing CoT effectiveness.

In summary, our key contributions are as follows: (1) We identify key factors that influence CoT's effectiveness on different reasoning tasks, including problem difficulty, information gain, and information flow. (2) We interpret the unfaithful CoT issue by jointly analyzing the information interaction among question, CoT, and answer. Based on experimental results, we demonstrate that the reason is that LLMs retrieve correct information (lost in the CoT) directly from the question when predicting answers. (3) As an application of our findings, we design a new method called QUIRE, which effectively improves the CoT's performance from the effectiveness (up to **2.4%** improvement) and faithfulness (up to **5.6%** improvement). This indicates that enhancing CoT faithfulness can lead to an improvement in CoT effectiveness. We release the source code in the attached software package.

2 Related Works

2.1 Chain-of-Thought Effectiveness

Since the emergence of CoT, a series of CoT-like approaches have further improved the model's reasoning accuracy through various prompt designs (Wang et al., 2023; Madaan et al., 2023b; Zhou et al., 2023). Recently, the emergence of reasoning models such as DeepSeek-R1 (DeepSeek-AI et al., 2025) and o1 (OpenAI, 2024) has once again proven that CoT is highly effective in solving complex reasoning tasks such as mathematics and coding (Qi et al., 2024; Snell et al., 2024; Zeng et al., 2024; Lightman et al., 2024). However, another series of works shows that the effectiveness of CoT has limitations (Wang et al., 2024b; Xu and Ma, 2024). They demonstrate that CoT brings only limited improvements in knowledge and commonsense reasoning tasks (Sprague et al., 2024), and may even harm the model's original performance (Li et al., 2024). Building on these studies, our work further investigates the key factors that control CoT's effectiveness across different tasks.

2.2 Chain-of-Thought Faithfulness

In model interpretability, faithfulness, defined as "accurately representing the reasoning process behind the model's decision", is important for evaluating the performance of natural language explanation (Ribeiro et al., 2016; Gilpin et al., 2018; Jacovi and Goldberg, 2020). With the emergence of CoT-like work, there has been increasing focus on measuring this characteristic within CoTs (Turpin et al., 2023; Lanham et al., 2023; Lyu et al., 2023). Some studies introduce counterfactual perturbations to questions and measure the change of answers (Atanasova et al., 2023; Turpin et al., 2023). Some other works use causal median analysis on CoTs and answers, calculating the treatment effect to represent the faithfulness (Bao et al., 2024; Paul et al., 2024). However, these works lack a comprehensive explanation and mitigation of unfaithful CoT, and this paper addresses this gap.

3 What Makes CoT Effective

In this section, we investigate what factors make the CoT effective in certain reasoning tasks. Specifically, we start with evaluating the final accuracy improvement of CoT on different tasks (§3.1). Then we study the impact of three different factors on the

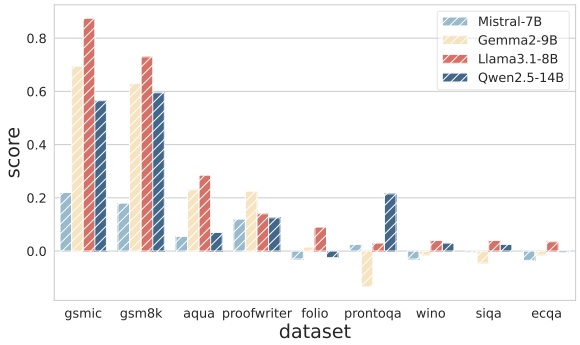
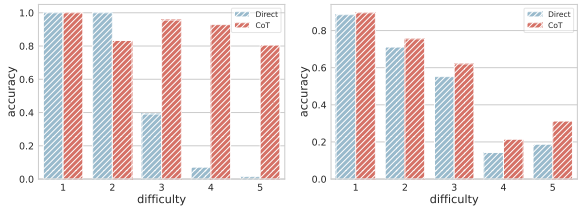


Figure 1: CoT improvement across different models and datasets, 'score' indicates the accuracy difference.



(a) GSM8k (b) WinoGrande

Figure 2: Performance on different problem difficulty levels with and without CoT prompting (Llama3.1-8B).

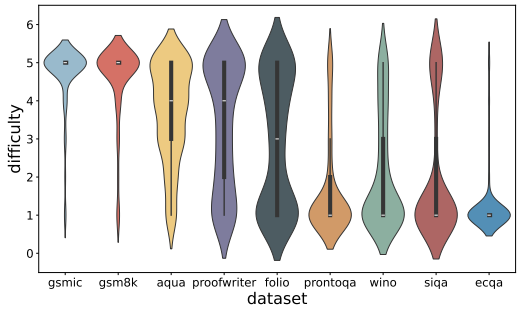


Figure 3: Difficulty distribution in different datasets.

final performance of CoT, including problem difficulty (§3.2), CoT information gain (§3.3), and the information flow between CoT and answer (§3.4).

3.1 Overall Performance

Experimental Setup We choose 9 representative datasets from various reasoning types for evaluation. Specifically, for mathematical reasoning, we choose GSMIC (Shi et al., 2023), GSM8K (GSM) (Cobbe et al., 2021) and AQuA (Ling et al., 2017). For logical reasoning, we choose ProofWriter (PW) (Tafjord et al., 2021), FOLIO (Han et al., 2024) and ProntoQA (PQA) (Saparov and He, 2023). For commonsense reasoning, we choose WinoGrande (WINO) (Sakaguchi et al., 2020), SocialIQA (SIQA) (Sap et al., 2019) and ECQA (Aggarwal et al., 2021). For models, due to the difficulty of deeply analyzing the internals of black-box models, we focus on analyzing white-box models and select four advanced white-box LLMs for the experiment, including Mistral-7B (Jiang et al., 2023), Gemma2-9B (Rivière et al., 2024a), Llama3.1-8B (Rivière et al., 2024b), and Qwen2.5-14B (Yang et al., 2024). For metrics, we define the effectiveness of CoT as the difference in accuracy when answering questions with and without CoT.

Main Results The main results of the evaluation experiment are illustrated in Figure 1, from which we can get that: **Among different reasoning tasks, CoT is most effective in mathematical reasoning, while least effective in commonsense reasoning tasks.** This conclusion forms the basis for the subsequent analysis in this section.

3.2 Problem Difficulty

Why is CoT more effective on certain task types? Reflecting on humans' reasoning process, the more difficult the problem, the more thinking time is required. Hence, we aim to explore whether this pattern can also be observed in LLMs: Is CoT more effective for harder problems?

Problem Difficulty Estimation Following former works, we classify the difficulty of questions based on the model's accuracy in answering them (Lightman et al., 2024; Setlur et al., 2024). Specifically, for each question, we sample 10 answers without CoT prompting and bin the average pass@1 rate across all models into five quantiles, each corresponding to increasing difficulty levels. For example, if the pass@1 rate is less than 0.1, the question is classified as the hardest level 5. Conversely, if the pass@1 rate is more than 0.8, the question is classified as the easiest level 1.

Performance across Difficulty Levels After classifying the question, we compare the effectiveness of CoT across different difficulty levels and illustrate part of the results in Figure 2 (more results in Appendix A). We can conclude that: **(Cl.1) CoT is more effective on challenging questions compared to simple ones.** For questions at low difficulty levels (e.g. level 1, level 2), CoT provides minimal accuracy improvement and even degrades performance. In contrast, CoT significantly increases reasoning accuracy across different tasks when the question is difficult (e.g. level 4, level 5).

Difficulty Distribution on Different Tasks We further evaluate the difficulty distribution of different tasks to explain the varying effectiveness. Figure 3 shows the results on Llama3.1-8B. In mathematical reasoning, most problems are of higher difficulty, whereas in commonsense reasoning, most problems are of lower difficulty. Combining Cl.1, we can infer that the CoT is more effective in mathematical reasoning since it has more difficult problems compared to other tasks. This provides an explanation for the effectiveness distribution shown in Figure 1 from the perspective of problem difficulty.

3.3 Information Gain

When we define the problem difficulty, we only consider the final result of LLM's reasoning. To conduct a more comprehensive analysis, we delve into the reasoning process and continue to identify key factors. In practice, a harder question tends to require more extra information to answer. Thus, here we focus on the information gain of CoT in the reasoning process.

Information Gain Definition In information theory, Information Gain (IG) quantifies the reduction in uncertainty of the target variable Y after adding a certain feature X:

$$IG(Y, X) = H(Y) - H(Y|X) \tag{1}$$

where $H(Y)$ represents the entropy of Y, and $H(Y|X)$ represents the conditional entropy of Y given the feature X. Similarly, in the context of LLM reasoning, given a question Q and a CoT C, we define the IG as follows:

$$\begin{aligned} IG(C, Q) &= H(C) - H(C|Q) \\ &= -\sum_{i=1}^{n} p(c_i|C_{i-1}) \log p(c_i|C_{i-1}) \\ &+ \sum_{i=1}^{n} p(c_i|C_{i-1}; Q) \log p(c_i|C_{i-1}; Q) \end{aligned} \tag{2}$$

Here, $p(\cdot)$ indicates the model's output probability, C_{i-1} is the first $i-1$ tokens of CoT, and n is the length of CoT. IG represents the degree to which the uncertainty of CoT is reduced by the question. The larger the IG, the more information CoT obtains from the question, hence the less additional information is provided by CoT itself.

Experiment and Analysis We conduct experiments across different datasets and demonstrate the results in Figure 4. Compared to Figure 1, this figure shows an opposite trend: mathematical reasoning has the lowest IG, while commonsense tasks exhibit the highest IG. This indicates that: **(Cl.2) CoT is more effective when it provides additional information not present in the problem itself** (e.g. gsmic, gsm8k, aqua). In contrast, when CoT is ineffective for performance improvement, it provides less extra information.

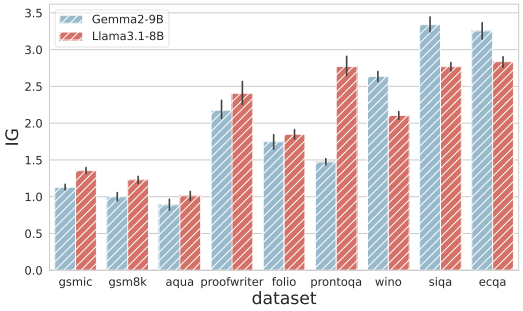


Figure 4: CoT information gain in different datasets.

3.4 Information Flow

In § 3.3, we primarily demonstrate the importance of additional information in CoT. However, does the way in which models utilize this information also affect the CoT effectiveness? To answer this question, we study the information flow between CoT and answers in this experiment.

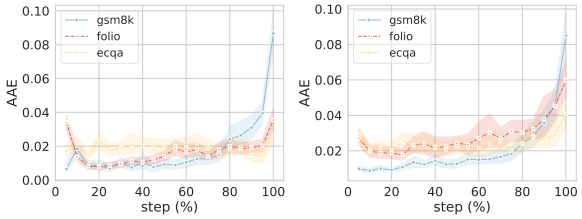
Information Tracing Method Following previous works (Wu et al., 2023; Wang et al., 2024a; Li et al., 2024), we employ integrated gradient attribution (IGA) (Sundararajan et al., 2017) as our measuring method to capture the information flow between CoT and answer. Specifically, we first compute importance $I_{n,m}$ of input token x_n to output token y_m as follows:

$$\begin{aligned} I(x_n, y_m) &= E(x_n) \int_{\alpha=0}^{1} \frac{\partial f(\alpha y_m)}{\partial E(x_n)} d\alpha \\ &\approx \frac{E(x_n)}{m} \sum_{k=1}^{m} \frac{\partial f(\frac{k}{m} y_m)}{\partial E(x_n)} \end{aligned} \tag{3}$$

where $f(\cdot)$ represents the model's output probability, $E(x_n)$ is the input word embedding of the token x_n and m is the number of approximation steps (we set it to 20). To reduce the interference from noise, we rescale the importance and get the attribution effect score between x_n and y_m:

$$AE(x_n, y_m) = \begin{cases} \frac{I(x_n, y_m)}{\max_{n'=1}^{N} I(x'_n, y_m)} & I(x_n, y_m) > 0 \\ 0 & otherwise \end{cases} \tag{4}$$

Here N is the last index of the input. Finally, we can measure the information flow between each



(a) Gemma2-9B (b) Llama3.1-8B

Figure 5: Information flow between the CoT and answer. 'Step' indicates sequential positions within the CoT, where 0 is the beginning and 100 is the end.

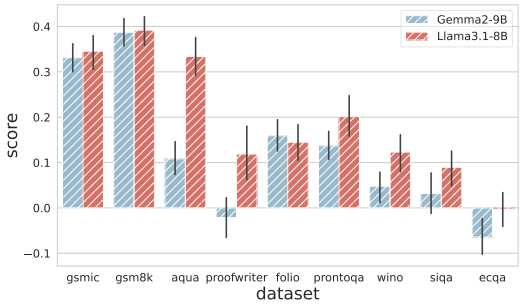


Figure 6: MIF score in different datasets.

token c of CoT and the answer A using the average attribution effect (AAE):

$$AAE(c, A) = \frac{1}{|A|} \sum_{a \in A} AE(c, a) \tag{5}$$

Since CoT is usually long, averaging over each token of CoT would result in a significant loss of information. Hence, we choose to average over A and analyze how the information flow changes throughout the CoT process using the AAE.

Information Flow Comparison We collect 200 CoT-answer pairs from three different datasets to calculate the AAE. Figure 5 shows the main results, from which we can get that: **(Cl.3) When information flow between CoT and the answer increases with the CoT process, the CoT tends to be effective.** As we can see from Figure 5, the curve of GSM8k exhibits the most significant upward trend, while ECQA remains the most stable, with the AAE showing little variation as the steps change. For tasks where CoT is highly effective (e.g. GSM8k), the influence of the CoT on the answer increases as the reasoning progresses. In contrast, for tasks that CoT is ineffective (e.g. ECQA), the influence of CoT on the answer does not significantly change as the reasoning progresses.

Monotonicity of Information Flow In the previous experiment, we identify the influence of AAE's increase by observing different curves. To quantitatively measure this increase, we define the monotonicity of information flow (MIF) as the Spearman correlation coefficient between the steps and the corresponding AAE values:

$$\begin{aligned} MIF(C, A) &= 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \\ &= 1 - \frac{6 \sum_{i=1}^{n} [n + 1 - i - R(AAE(c_i, A))]^2}{n(n^2-1)} \end{aligned} \tag{6}$$

where n is the length of CoT and $R(\cdot)$ is the ranking of the value. In the implementation, we merge adjacent tokens and calculate their average AAE, thereby reducing noise interference. The experimental results on Gemma2-9B and Llama3.1-8B are presented in Figure 6, from which we can get that: **The higher the monotonicity of the information transfer between CoT and the answer, the more effective the CoT becomes.** This further demonstrates the validity of Cl.3.

C. → A.	GSM	AQuA	PW	PQA	WINO	SIQA
✓ → ✓	41	25	14	27	34	40
✓ → ✗	0	0	0	0	1	0
✗ → ✓	1	1	7	17	1	0
✗ → ✗	8	24	29	6	14	10

Table 1: Inconsistency statistics between the CoT (C.) and the answer (A.) on Llama3.1-8B.

4 What Makes CoT Unfaithful

In this section, we aim to analyze the CoT from the faithfulness perspective. Concretely, we first identify the unfaithfulness problem in different tasks (§4.1). Next, we analyze the issue by examining the information interaction among the three key components of reasoning (as illustrated in Figure 7), including question and CoT (§4.2), CoT and answer (§4.3), question and answer (§4.4).

4.1 CoT Faithfulness Evaluation

Following previous works (Bao et al., 2024; Lyu et al., 2023), we evaluate the faithfulness of CoT by measuring the consistency between the CoT and the answer. If an incorrect CoT induces a correct answer or a correct CoT induces a wrong answer, it is seen as an unfaithful CoT (see Figure 7 for example). We manually evaluate the correctness of 50 CoT-answer pairs from six datasets and compare inconsistency ratios in them. The main results on Llama3.1-8B are illustrated in Table 1 (results on other models are presented in Appendix B). We

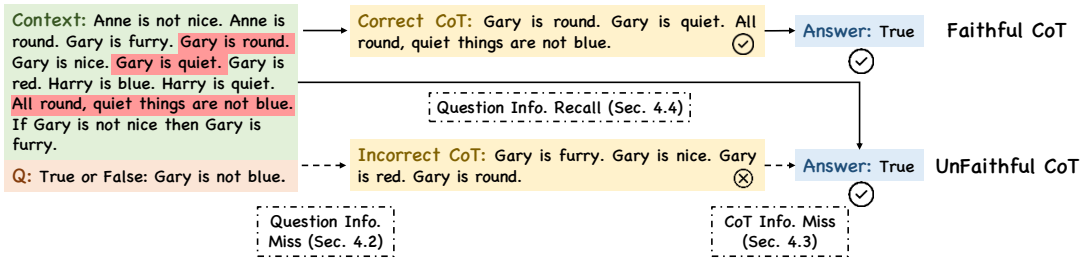
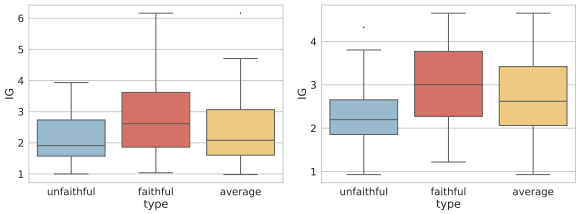
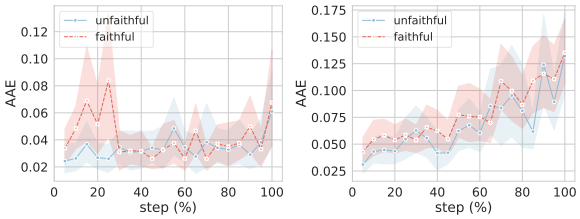


Figure 7: An interpretation of unfaithful CoT issues, where statements in red are correct information for reasoning.



(a) ProofWriter (b) ProntoQA

Figure 8: Comparison of information transfer between questions and CoTs under three settings.



(a) ProofWriter (b) ProntoQA

Figure 9: Comparison of information transfer between CoTs and answers on Llama3.1-8B.

can conclude that: **The logical reasoning tasks have more unfaithful CoT issues.** Compared to other datasets, the proportion of inconsistencies is higher in logical reasoning (7/50 in PW and 17/50 in PQA) and mainly consists of wrong CoTs leading to correct answers. Our research focuses on interpreting these unfaithful issues within logical reasoning datasets in the following sections.

4.2 Question to CoT: Unfaithful CoT misses correct information from context

We seek to explore why CoTs lack such correct information in unfaithful cases. Since CoTs are generated based on the question, we hypothesize that it is due to the lack of information from the context of the question. To demonstrate it, we use IG (see Eq.2) to compare the information interaction between questions and CoTs.

Experimental Setup We experiment with three settings: 'unfaithful', 'faithful', and 'average'. For 'unfaithful', we select all of the unfaithful samples, calculating $IG(Q, C)$. For 'faithful', we select samples where both CoT and the answer are correct (see Figure 7 for examples). For 'average', we calculate IG on all questions. We collect 200 samples from ProofWriter and ProntoQA, comparing the IG distribution under different settings.

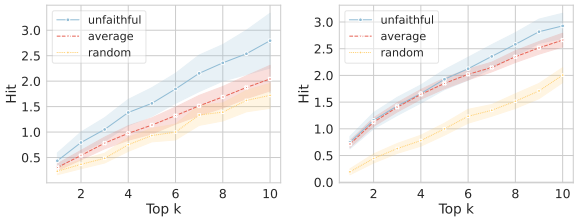
Experimental Results Figure 8 presents our results (we present more experiments in Appendix C). We can get that: **(Cl.4) Unfaithful CoT misses correct information from the context.** In both figures, the IG under the 'unfaithful' setting is lower than the other two settings. This indicates that CoT gets less information from the context when an unfaithful issue occurs. As an example, in unfaithful CoT of Figure 7, the incorrect CoT does not contain the statement "*Gary is quiet*" or "*All round, quiet things are not blue*" in the question.

4.3 CoT to Answer: Unfaithful CoT has less information transfer to answers

Since unfaithful CoT lacks the correct information needed for reasoning, why can the final prediction still be correct? To answer it, we investigate the information transfer between CoT and the answer.

Experimental Setup We use the AAE from the Eq.5 to measure the amount of information transferred between the two. Following the experiment in §4.2, we experiment under "unfaithful" and "faithful" settings, comparing AAE values on Llama3.1-8B across different datasets.

Experimental Results The main results of the experiments are demonstrated in Figure 9. In both figures, the AAE for the 'faithful' setting (in red) is higher than that for the 'unfaithful' setting (in blue). Therefore, we have: **(Cl. 5) Unfaithful CoT has less information interaction with the answer compared to the correct one.**



(a) ProofWriter (b) ProntoQA

Figure 10: Comparison of correct recall counts.

4.4 Question to Answer: Answer can recall correct information from context

While the answer misses key information from the CoT, how can the final prediction still be correct? We hypothesize that LLMs can recall the missing information when generating the answer and design experiments to demonstrate it.

Experimental Setup We rank each statement in the context by its AAE score to the answer $AAE(S, A)$ (S is a statement in the question) and observe whether the top-ranked statements include the correct statement missing in CoT (e.g. "*Gary is quiet*" in Figure 7). For comparison, we conduct experiments under three settings: unfaithful (unfaithful CoT with AAE recall), average (all CoT with AAE recall), and random (unfaithful CoT with random recall).

Experimental Results Figure 10 demonstrates our results on Llama3.1-8B (results on more models in Appendix D), from which we conclude that: **(Cl. 6) When unfaithful CoT issues occur, LLMs can recall missing correct information from the context during the answer prediction.** For all datasets and models, when the unfaithful CoT issue occurs, more missing statements get the top-k highest AAE scores from the answer compared to other settings. These statements have a strong information interaction with the answer, compensating for the lack of relevant statements in the CoT, thereby contributing to the correct answer prediction.

5 From Unfaithful CoT to Effective CoT

Since we analyze the CoT from two different perspectives in the former experiments, what is the relationship between them? In this section, we demonstrate that mitigating the unfaithful issue can lead to improvements in final performance. In other words, the faithfulness of CoT (§4) is a key factor in influencing the CoT effectiveness (§3).

5.1 Our Method

Based on findings in §4, we propose a new method called **QU**estion **I**nformation **R**ecall and **E**nhancement (QUIRE) to mitigate the unfaithful CoT issue. The main framework of it is illustrated in Figure 11, which includes two components:

AAE Recall As mentioned in Cl.6, when unfaithful issues occur, LLMs maintain a strong causal relevance with the correct statement in the context during the answer prediction. Thus, here we first generate a raw answer A with the SC method, then recall extra information by selecting the top-k context statements with the highest $AAE(S, A)$ (as marked with red in Figure 11). After recalling extra information, we incorporate these statements as additional hints into the input prompt, enabling the model to pay more attention to this information during the CoT generation.

IG Vote Through the former step, we get multiple information-enhanced CoTs (here we can also integrate the SC technique to further improve the performance). However, since our recall method also introduces noisy hints, there may exist incorrect statements in some of these CoTs (e.g. Hint 1 in Figure 11). To reduce their interference, according to Cl.4, we rate these CoTs based on $IG(Q, C)$. A higher IG indicates that more information in CoT is derived from the question, which means the CoT contains fewer hallucinated statements. After calculation, we use these scores as the weight for SC to vote and select the final answer.

5.2 Main Experimental Setup

Datasets Since all analyses in §4 are conducted on ProofWriter (Tafjord et al., 2021) ProntoQA (Saparov and He, 2023), we continue to evaluate our method on them. For the test set, we sample 500 questions from the former and 400 questions from the latter.

Metrics In form sections, we analyze the CoT performance from two aspects. Therefore, our evaluation cannot solely consider the result performance but should also assess the quality of the CoT to avoid unfaithful reasoning. Therefore, in addition to accuracy (Acc), we use the following two metrics: **(1) BertScore (BS):** Given a golden rationale, the generated CoT should recall as much information from it as possible, hence, we use the BertScore (Zhang et al., 2020) as one of our metrics. **(2) Faithful BertScore (FBS):** From the per-

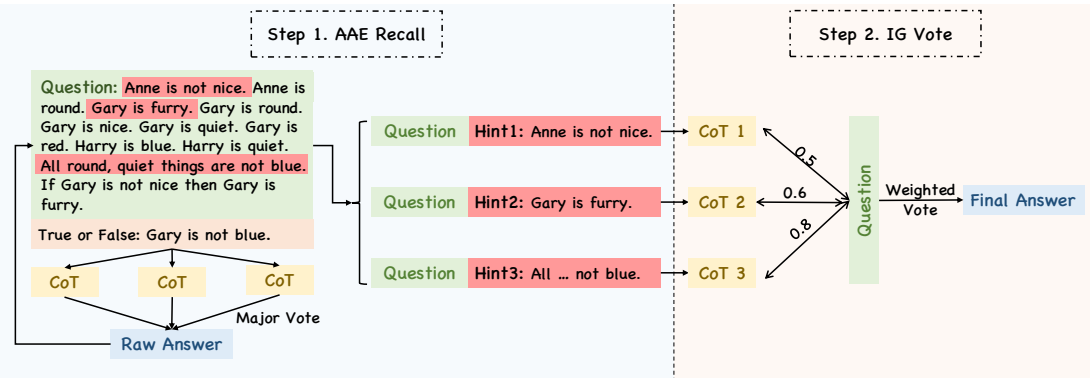


Figure 11: The main process of our QUIRE method, where the statement in red is the recalled information.

Method	ProofWriter			ProntoQA		
	Acc	BS	FBS	Acc	BS	FBS
CoT	59.2	64.9	55.7	86.8	86.1	78.0
SC	60.6	65.0	57.8	93.2	87.5	83.6
LtM	54.0	60.4	56.4	90.0	77.3	72.6
SR	51.6	65.9	53.4	88.5	91.5	84.5
Ours	**63.0**	**66.6**	**58.0**	**95.0**	**92.7**	**89.2**
- AAE Recall	60.2	65.1	57.0	95.0	87.5	84.6
- IG Vote	62.8	64.1	56.6	94.3	87.0	83.4

Table 2: Results of our main experiment, the best results are highlighted in **bold**.

spective of faithfulness, correct answers should be accompanied by high-quality CoTs, and incorrect results should correspond to CoTs of poorer quality. Thus, we define the FBS to measure faithfulness as below:

$$FBS = \frac{1}{n}\sum_{i=1}^{n}[\eta(a_i)BS(c_i, g_i) + (1-\eta(a_i))(1-BS(c_i, g_i))] \quad (7)$$

where c_i, a_i, g_i represent the generated CoTs, answers and golden rationales, n denotes the sample count. If a_i is correct, $\eta(a_i) = 1$, else $\eta(a_i) = 0$.

Baselines For baselines, we select representative methods that enhance LLMs' reasoning performances, including: **Chain-of-Thought (CoT)** (Wei et al., 2022), **Self-Consistency (SC)** (Wang et al., 2023), **Least-to-Most (LtM)** (Zhou et al., 2023), **Self-Refine (SR)** (Madaan et al., 2023b). Additionally, we also set up ablation experiments (-AAE Recall and -IG Vote) to verify the effectiveness of each component in our method. Implementation details can be found in Appendix E.

5.3 Main Experimental Results

The results of our main experiment on Llama3.1-8B are demonstrated in Table 2 (additional results in Appendix F), which demonstrates that: **(1) Our method effectively mitigates the unfaithful CoT issues.** On both BS and FBS, our method achieves the highest performance, improving up to 5.6% faithfulness (i.e. FBS) on ProntoQA. Besides, from the results of the ablation study, we can see both modules make contributions to enhancing the CoT faithfulness. Given that our method is an application derived from the analytical conclusions, its superior performance can also substantiate the correctness of our earlier findings. **(2) Improvements in faithfulness can also lead to enhancements in CoT's effectiveness.** Although our method is based on the conclusions from §4 to optimize the unfaithful CoT issue, the CoT effectiveness (Acc) also improved (up to **2.4%** on ProofWriter), indicating that the former is a significant factor influencing the latter. Through our method, we can boost the CoT's performance from both effectiveness and faithfulness.

6 Conclusion

In this paper, we focus on analyzing the CoT performance in reasoning tasks. Specifically, we identify the factors influencing CoT effectiveness and interpret the mechanism behind CoT unfaithfulness. For the former, we conduct extensive experiments to demonstrate that question difficulty, information gain, and information flow all contribute to CoT's performance improvement. For the latter, we capture the information transfer among questions, CoTs, and answers in the reasoning process. The experimental results indicate that the information recall mechanism during answer predictions leads to unfaithful CoT issues. At last, we design the QUIRE method as a preliminary application of our findings, which significantly improves CoT performances from both perspectives.

Limitations

Although our work conducts an in-depth analysis and proposes mitigation strategies for improving CoT performance, it has several limitations. Firstly, due to the inability to access gradient information inside models like GPT-4, our analysis is limited to open-source LLMs. Secondly, although we have empirically demonstrated that improvements in faithfulness can lead to performance enhancements, there is still a lack of corresponding theoretical proof to support this conclusion. We leave the CoT effectiveness analysis of black-box LLMs and further theoretical proof for our future work.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3050–3065. Association for Computational Linguistics.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 283–294. Association for Computational Linguistics.

Guangsheng Bao, Hongbo Zhang, Linyi Yang, Cunxiang Wang, and Yue Zhang. 2024. Llms with chain-of-thought are non-causal reasoners. *CoRR*, abs/2402.16048.

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.

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, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, pages 80–89. IEEE.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander R. Fabbri, Wojciech Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. 2024. FOLIO: natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Lan-*

guage Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 22017–22031. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4198–4205. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, 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. *CoRR*, abs/2307.13702.

Jiachun Li, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. 2024. Focus on your question! interpreting and mitigating toxic cot problems in commonsense reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9206–9230. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 305–329. Association for Computational Linguistics.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023a. What makes chain-of-thought prompting effective? A counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1448–1535. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023b. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024. Introducing openai o1 preview. Accessed: 2025-01-24.

Letitia Parcalabescu and Anette Frank. 2023. On measuring faithfulness of natural language explanations. *arXiv preprint arXiv:2311.07466*.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *CoRR*, abs/2402.13950.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *CoRR*, abs/2408.06195.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock,

Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024a. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024b. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4462–4472. Association for Computational Linguistics.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. Rewarding progress: Scaling automated process verifiers for LLM reasoning. *CoRR*, abs/2410.08146.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *CoRR*, abs/2408.03314.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *CoRR*, abs/2409.12183.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3621–3634. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. 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 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. 2024a. Gradient based feature attribution in explainable AI: A technical review. *CoRR*, abs/2403.10415.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024b. Chain-of-probe: Examing the necessity and accuracy of cot step-by-step. *CoRR*, abs/2406.16144.

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 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. 2023. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. *CoRR*, abs/2310.00492.

Nan Xu and Xuezhe Ma. 2024. Llm the genius paradox: A linguistic and math expert's struggle with simple word-based counting problems. *arXiv preprint arXiv:2410.14166*.

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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. 2024. Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective. *CoRR*, abs/2412.14135.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

A Additional Experiments across Different Difficulty Levels

In the main text, due to space constraints, we only presented results on GSM8k and WinoGrande, here we show more results on other datasets and models in Figure 12, 13, 14, 15. Besides, we also show the difficulty distribution on more models in Figure 16 and 17. These results are consistent with our conclusions in Cl.1.

B Details and Additional Experiments on Faithfulness Evaluation

To demonstrate the widespread existence of unfaithful issues in logical tasks, we present the evaluation results on Llama2-13B in Table 3.

C Additional Experiments on Question to CoT Information Analysis

In addition to the main experiments in §4.2, here we conduct another experiment to further demonstrate our conclusion in Cl.4. Specifically, we experiment with three settings: 'miss', 'hit', and 'avg'. For 'miss', we select the context statements that are present in the golden CoT (as provided in the dataset) but not in the generated CoT, calculating their $AAE(Q, C)$ scores to CoT. For 'hit', we collect the statements present in the generated CoT and compute the corresponding $AAE(Q, C)$. As for 'avg', we calculate the $AAE(Q, C)$ between the whole context and CoT. We compare the distribution of the above three AAE scores on ProofWriter and ProntoQA (100 samples each) and illustrate the results in Figure 18, 19. Across all figures, the AAE for the 'hit' setting is higher than that for the 'miss' setting. Thus, compared to the question information present in the CoT, this missing information gets less attention from the model during the CoT generation. Besides, the score difference between the 'hit' and the 'avg' is also large, which means that the included context statements have a stronger information interaction with the CoT. The model tends to copy this attended information into the CoT. Therefore, the results are consistent with our findings in §4.2.

D Additional Experiments on Question to Answer Information Analysis

To demonstrate the generalizability of our conclusions in §4.4, we repeat the experiments on two more models and present the result in Figure 20 and 21 (here we sample 100 questions from ProntoQA and ProofWriter). The results are consistent with Cl.6.

E Implementation Details of the Main Experiment

Here we provide a detailed account of the implementation specifics from the main experiments in §5. For SC, we generate 3 samples for each question since our method is also set to 3 paths. For our method, we recall top-3 statements in AAE recall and generate one CoT for each enhanced prompt. We release all the prompts we use in the attached software package.

F Additional Experiments on the Main Experiment

We also repeat the experiments on Gemma2-9B and report the results in Table 4, which demonstrates the effectiveness of our method.

C. → A.	GSM	AQuA	PW	PQA	WINO	SIQA
✓ → ✓	11	3	13	22	16	31
✓ → ✗	0	0	5	1	7	0
✗ → ✓	0	7	23	19	6	0
✗ → ✗	39	40	9	8	21	19

Table 3: Inconsistency statistics between CoTs and answers on Llama2-13B.

Method	**ProofWriter**			**ProntoQA**		
	Acc	BS	FBS	Acc	BS	FBS
CoT	65.0	56.6	52.9	77.0	62.7	57.7
SC	31.0	54.0	50.3	81.0	64.5	60.5
LtM	55.0	55.4	51.8	90.0	71.0	67.6
SR	18.5	43.1	58.6	56.5	45.3	51.9
Ours	**65.0**	**60.7**	**56.3**	**92.5**	**71.2**	**69.5**
- AAE Recall	27.5	54.2	50.3	89.0	64.5	61.9
- IG Vote	58.5	57.8	52.8	74.5	65.9	60.6

Table 4: Results of our main experiment on Gemma2-9B, the best results are highlighted in **bold**.

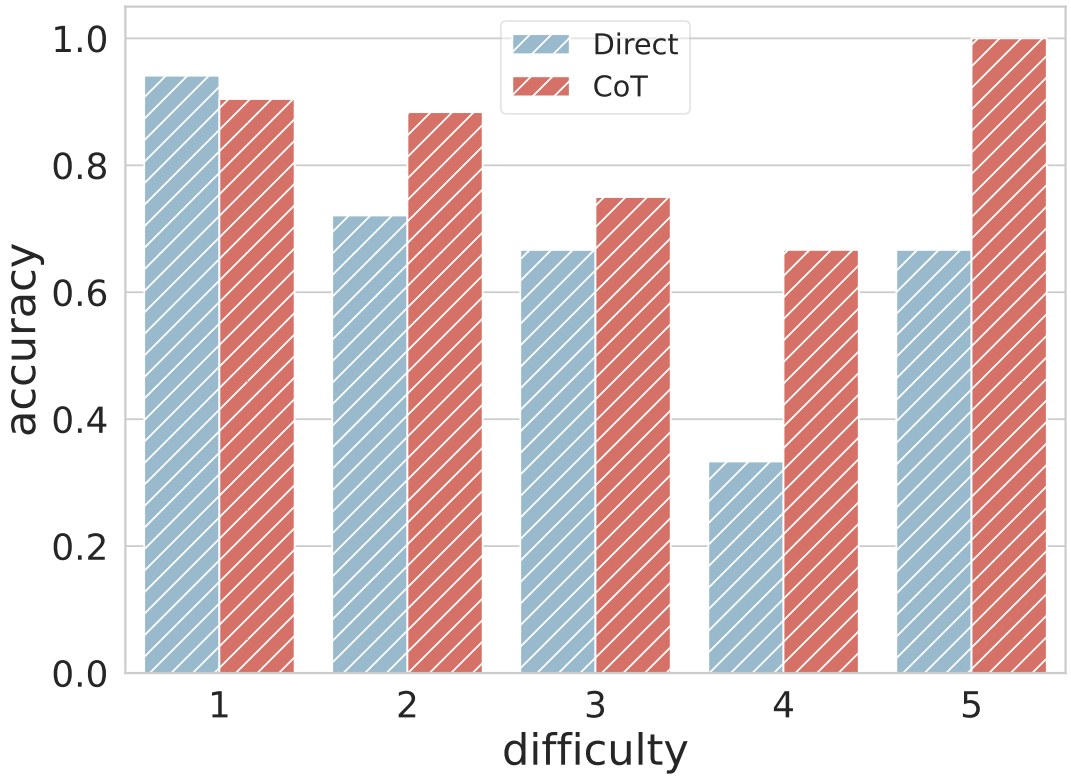


Figure 12: Performance on different problem difficulty levels with and without CoT prompting (Llama3.1-8B on ProntoQA).

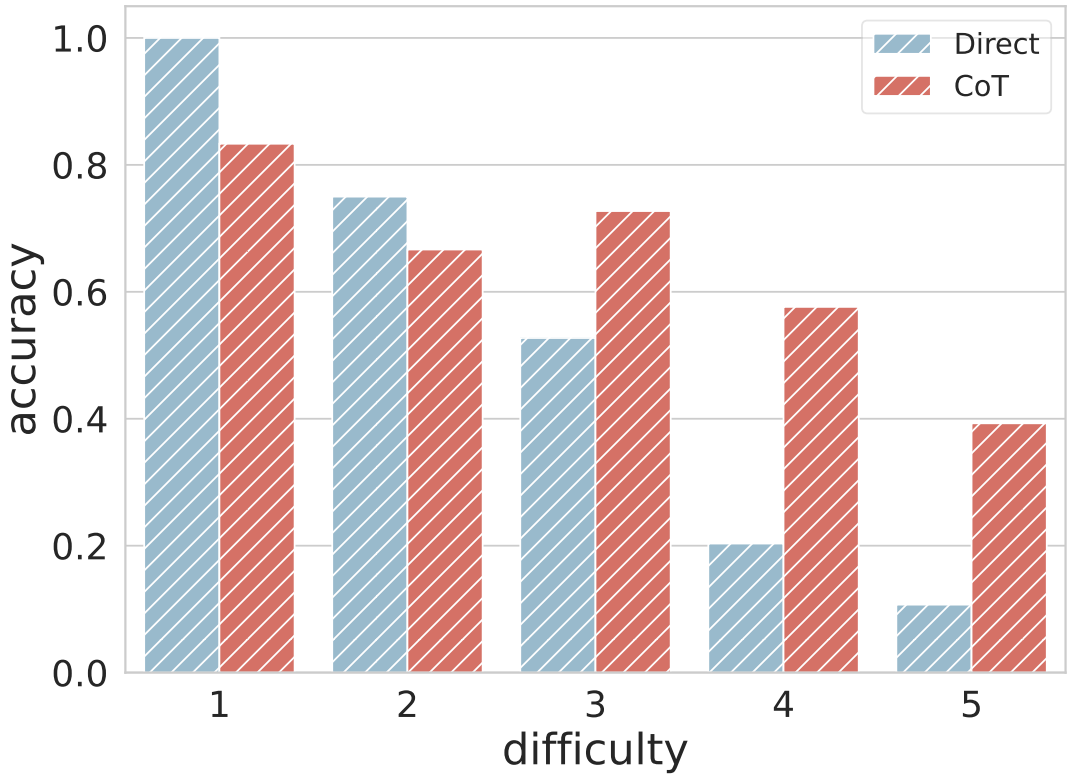


Figure 13: Performance on different problem difficulty levels with and without CoT prompting (Gemma2-9B on AQuA).

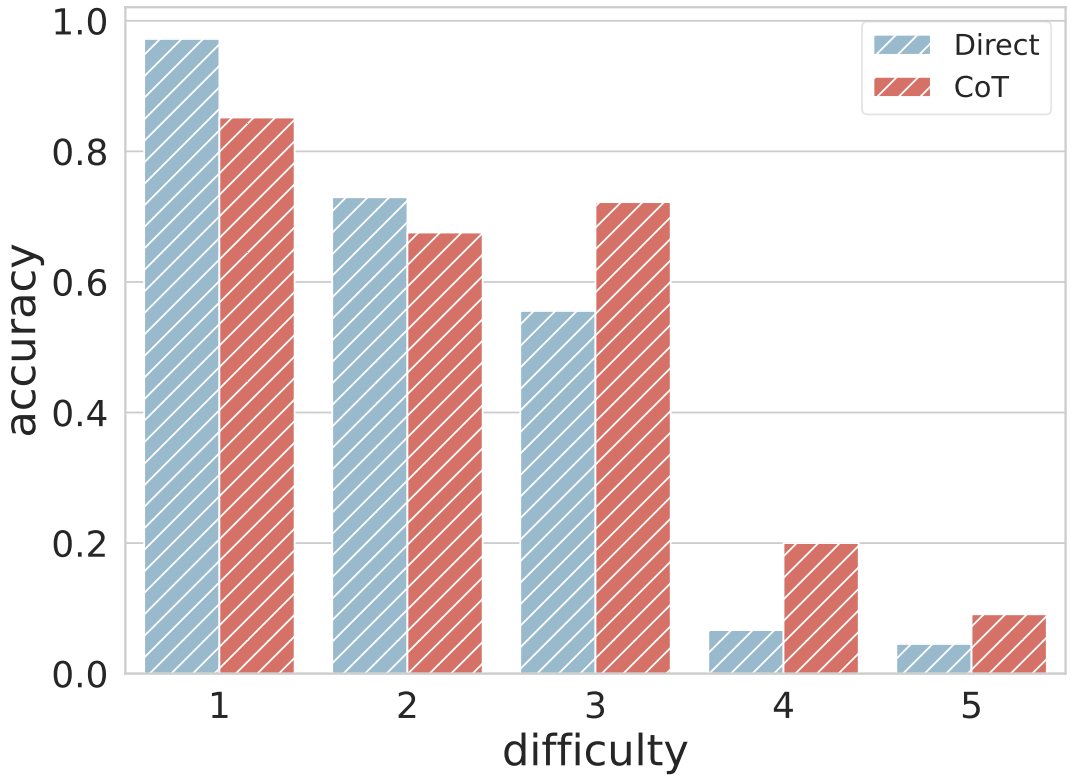


Figure 14: Performance on different problem difficulty levels with and without CoT prompting (Gemma2-9B on SIQA).

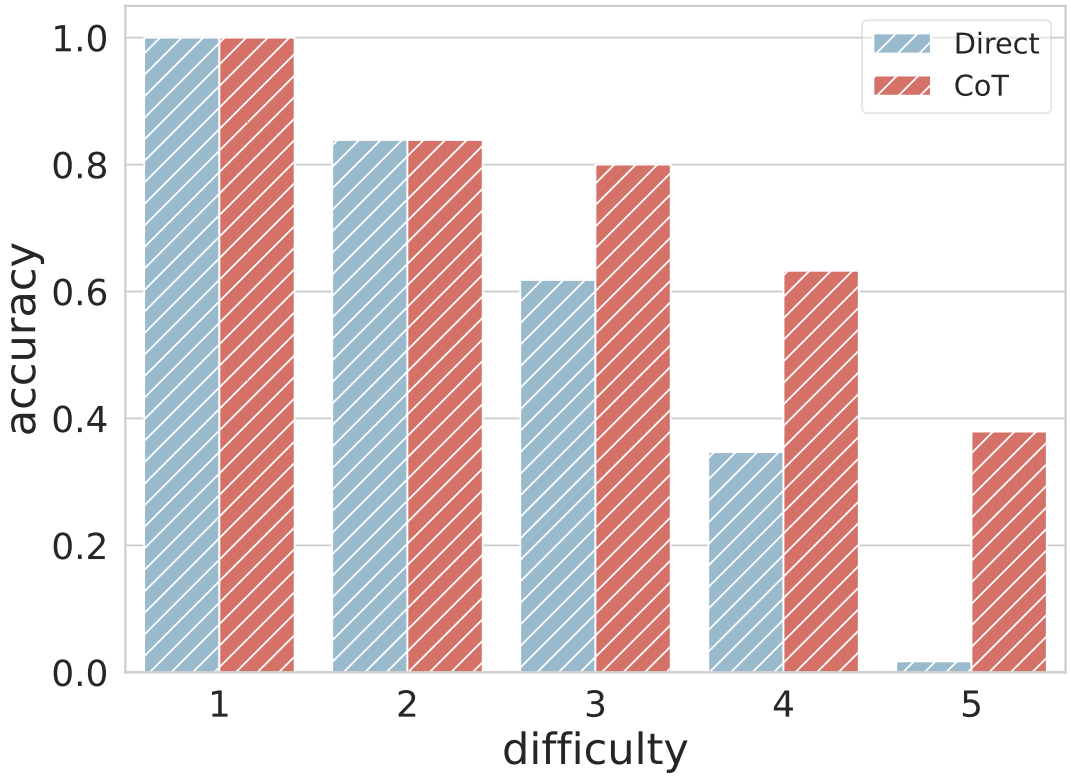


Figure 15: Performance on different problem difficulty levels with and without CoT prompting (Gemma2-9B on ProofWriter).

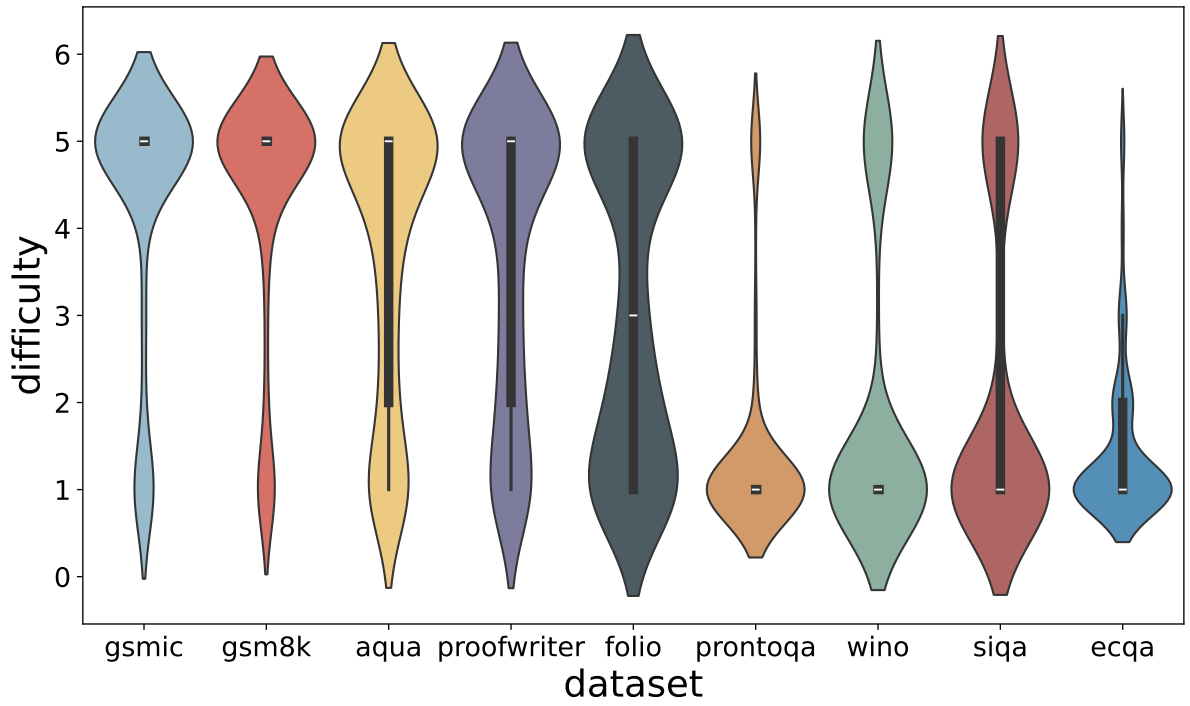


Figure 16: Difficulty distribution in different datasets on Gemma2-9B.

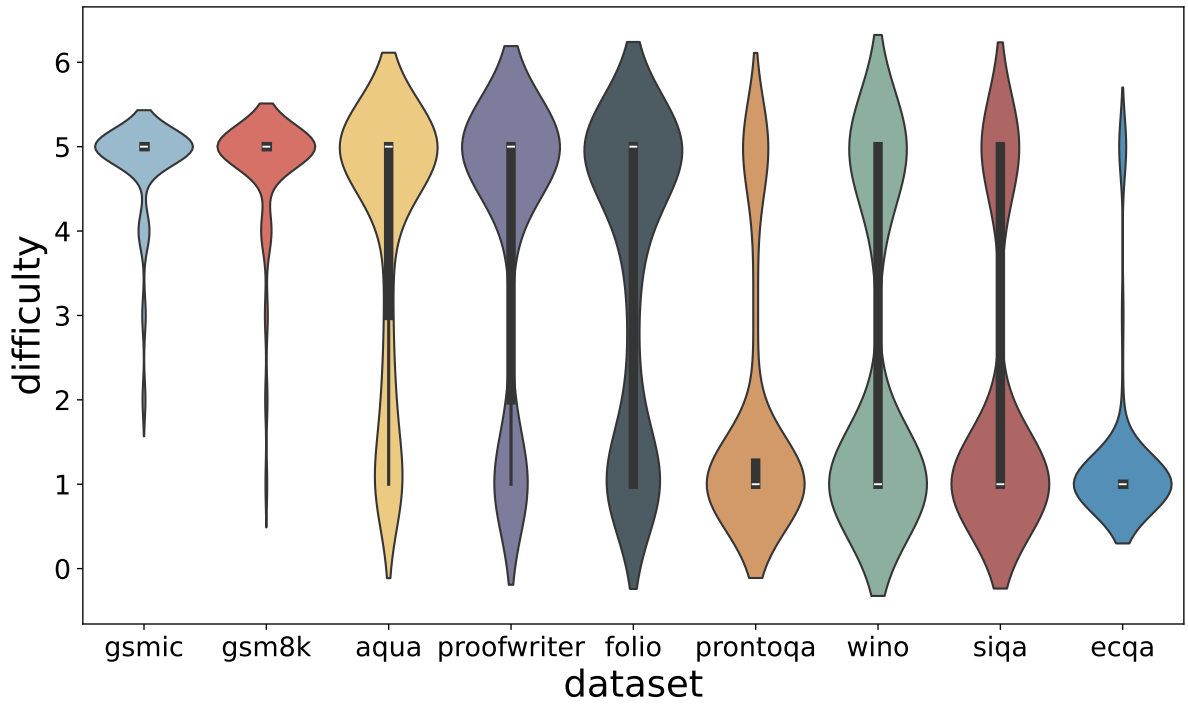
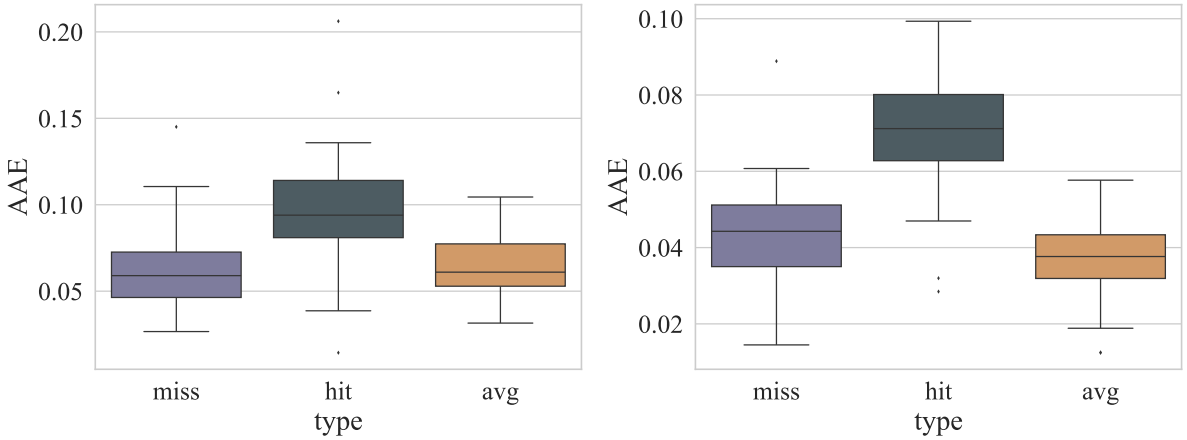


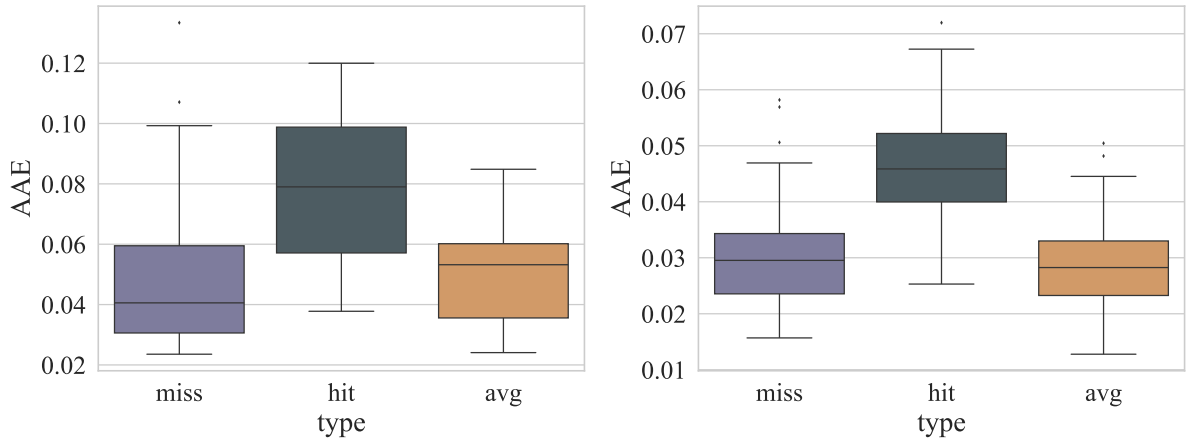
Figure 17: Difficulty distribution in different datasets on Mistral-7B.



(a) ProofWriter

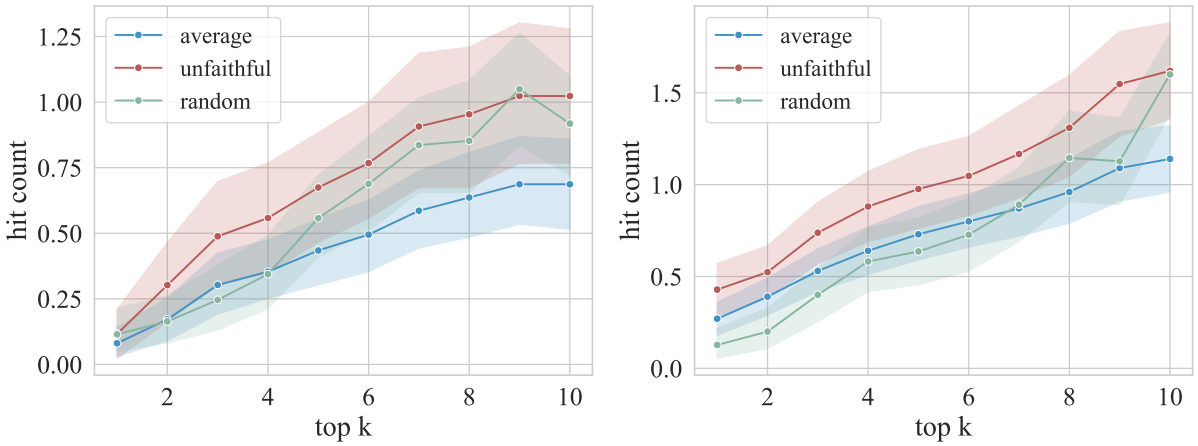
(b) ProntoQA

Figure 18: Comparison of information interaction between contexts and CoTs under three settings (Llama2-13B).



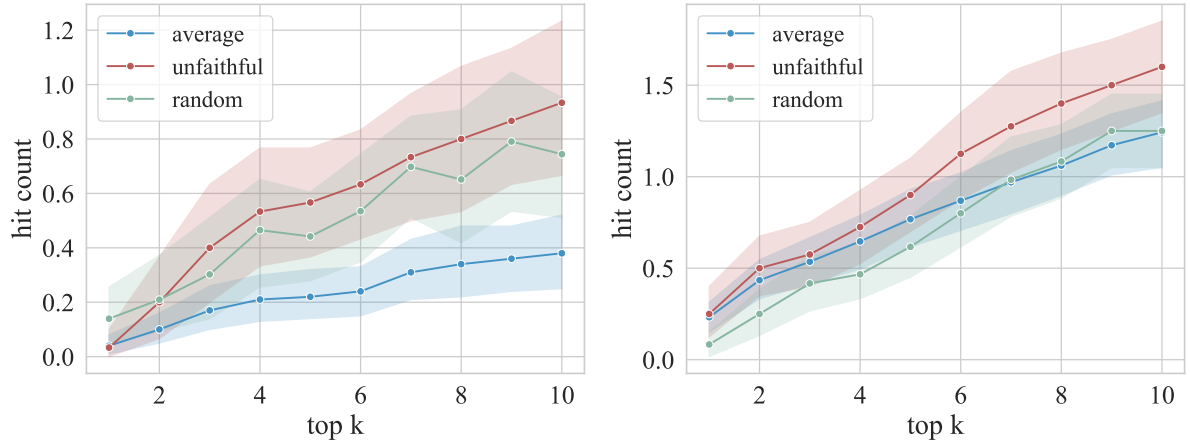
(a) ProofWriter (b) ProntoQA

Figure 19: Comparison of information interaction between contexts and CoTs under three settings (Mistral-7B).



(a) ProofWriter (b) ProntoQA

Figure 20: Comparison of correct statements recall counts (Llama2-13B).



(a) ProofWriter (b) ProntoQA

Figure 21: Comparison of correct statements recall counts (Mistral-7B).