
Causal Sufficiency and Necessity Improves Chain-of-Thought Reasoning

Xiangning Yu^{1,6*}, Zhuohan Wang^{2*}, Linyi Yang³, Haoxuan Li⁴,

Anjie Liu⁵, Xiao Xue^{1†}, Jun Wang³, Mengyue Yang^{6‡}

¹Tianjin University ²City University of Hong Kong

³University College London ⁴Peking University

⁵The Hong Kong University of Science and Technology (Guangzhou) ⁶University of Bristol

Abstract

Chain-of-Thought (CoT) prompting plays an indispensable role in endowing large language models (LLMs) with complex reasoning capabilities. However, CoT currently faces two fundamental challenges: (1) Sufficiency, which ensures that the generated intermediate inference steps comprehensively cover and substantiate the final conclusion; and (2) Necessity, which identifies the inference steps that are truly indispensable for the soundness of the resulting answer. We propose a causal framework that characterizes CoT reasoning through the dual lenses of sufficiency and necessity. Incorporating causal Probability of Sufficiency and Necessity allows us not only to determine which steps are logically sufficient or necessary to the prediction outcome, but also to quantify their actual influence on the final reasoning outcome under different intervention scenarios, thereby enabling the automated addition of missing steps and the pruning of redundant ones. Extensive experimental results on various mathematical and commonsense reasoning benchmarks confirm substantial improvements in reasoning efficiency and reduced token usage without sacrificing accuracy. Our work provides a promising direction for improving LLM reasoning performance and cost-effectiveness. The code is available at: <https://github.com/yxn9191/causalmath>.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive advancements in complex reasoning tasks, significantly attributed to the adoption of Chain-of-Thought (CoT). CoT prompting guides models to generate intermediate reasoning steps, thereby enhancing performance in areas such as arithmetic problem-solving and commonsense reasoning [43, 17, 8]. Despite these improvements, CoT reasoning faces two fundamental challenges: (i) **Sufficiency**: ensuring that the generated intermediate steps comprehensively support the conclusion [50, 3, 33], and (ii) **Necessity**: identifying which steps are indispensable for the soundness of the final answer [7, 54]. Figure 1a illustrates three common reasoning patterns frequently observed in LLMs, exemplified here using a GSM-8k [10] question: (1) *Sufficient but Unnecessary*, where redundant steps reduce reasoning efficiency; (2) *Necessary but Insufficient*, in which incomplete reasoning fails to reach the correct answer; and (3) *Sufficient and Necessary*, the ideal case that balances correctness and conciseness. These examples highlight the impact of reasoning inefficiencies—especially “overthinking”, where unnecessary steps may hinder rather than help model performance.

*Equal contribution.

†Corresponding authors.

‡Project leader

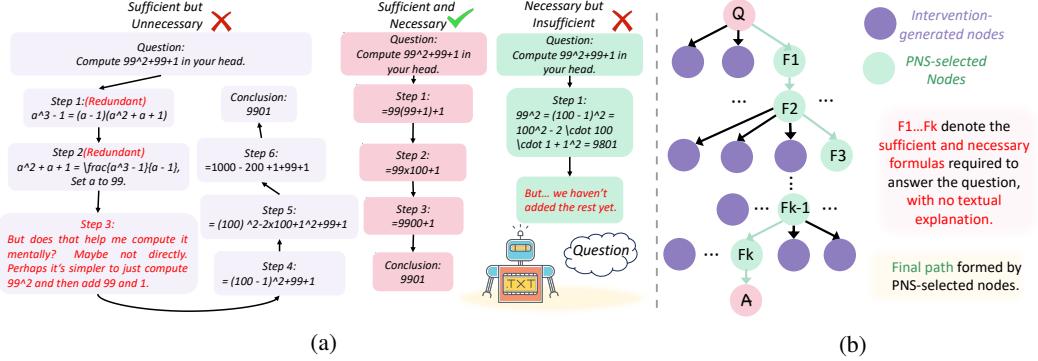


Figure 1: (a) Illustration of three reasoning types—*Sufficient but Unnecessary*, *Necessary but Insufficient*, and *Sufficient and Necessary*—based on actual model-generated responses to a GSM-8k question: “Compute $99^2 + 99 + 1$ in your head.” (b) Path selection process using our method. Purple nodes denote CoT steps obtained through causal intervention (rollout), while green nodes indicate the minimal steps satisfying both sufficiency and necessity.

Recent research on Chain-of-Thought (CoT) reasoning has addressed **Sufficiency** by introducing strategies such as self-consistency decoding [59] and iterative refinement methods like Self-Refine [40], aiming to ensure intermediate steps comprehensively support final answers [27, 48, 17]. Concurrently, efforts targeting **Necessity** have developed pruning techniques, such as addressing the “overthinking” by reducing the token length [7, 36]. Chain-of-Draft prompting [67] and identify critical reasoning steps [13], to reduce redundancy in reasoning paths [42, 62, 52, 45]. However, none have utilized rigorous mathematical analyses based on sufficient and necessary conditions [46] to evaluate and prune reasoning paths. These methods predominantly rely on correlation-based metrics (e.g., attention weights, likelihood scores, or ablation accuracy), which may misleadingly associate frequent or prominent steps with correctness without verifying true causal impact [4]. Consequently, correlation alone is insufficient for reliably distinguishing genuinely necessary or sufficient reasoning steps, highlighting the need for causal frameworks to rigorously assess their logical contributions.

To jointly address the sufficiency and necessity of reasoning steps while ensuring logical and causal soundness, we introduce the concept of causal Probability of Necessity and Sufficiency (PNS) and redefine it for CoT reasoning framework. We theoretically analyse the identifiability of PNS in CoT. Based on the identifiability results, we develop a PNS-based evaluation algorithm to systematically reconstruct reasoning sequences by causal intervention (rollout) (shown in Figure 1b). Using this algorithm, we effectively reconstruct CoT responses from training data that explicitly meet causal sufficiency and necessity criteria, thus eliminating redundant steps without compromising—and potentially enhancing—answer accuracy. The reconstructed reasoning CoT then serve as causally-informed demonstrations, enabling LLMs to acquire causal reasoning capabilities via in-context learning and fine-tuning to improve the efficiency without sacrificing the accuracy. Empirical evaluations on mathematical reasoning benchmarks—including GSM-8k [10], MATH-500 [25], and AIME [44], as well as the CommonsenseQA [53] dataset—confirm that our approach significantly reduces reasoning redundancy while maintaining or improving prediction accuracy.

Our main contributions are as follows:

1. We propose a conceptual integration of Probability of Necessary and Sufficient causation (PNS) into CoT reasoning.
2. We introduce a novel bi-level optimization framework based on PNS for systematically constructing efficient and accurate CoT reasoning sequences.
3. We empirically validate our approach across diverse reasoning tasks, demonstrating substantial improvements in both efficiency and accuracy through optimized CoT traces used for in-context learning and supervised fine-tuning.

2 Related Work

Reasoning Sufficiency Enhancement via CoT Optimization. CoT reasoning [61] has significantly improved LLM performance on complex tasks, inspiring variants such as Tree-of-Thought [76], Graph-of-Thought [6], and DOTS [78]. Further developments include multimodal extensions [65], latent variable formulations [64], and dynamic memory usage [73]. Others enhance reasoning via self-correction [47], counterfactual fine-tuning [26], or prompt design [18]. Despite strong performance, many methods suffer from unnecessary verbosity, inefficient computation, or overthinking [62, 7]. Beyond textual reasoning, task-grounded reasoning frameworks such as ChessGPT [20] bridge policy learning and language modeling, revealing that reasoning sufficiency can also be optimized in structured decision domains.

Reasoning CoT Redundancy. Recent work targets CoT redundancy by compressing reasoning traces (e.g., C3oT [32], CoT-Valve [39], CCoT [9]), pruning superfluous steps [36], or using token-budget-aware reasoning [22]. Training-free approaches such as Kimi [54] and external thought injection [35] further optimize reasoning cost. SPIRIT [13] leverages perplexity to identify key reasoning steps, balancing accuracy and efficiency in both few-shot and fine-tuned CoT settings, while also generalizing well across models. Ton et al. [57] use conditional mutual information to quantify each step’s contribution to the final answer, revealing failure patterns without requiring intermediate supervision. However, existing methods often prioritize brevity or representational efficiency, without explicitly enforcing causal sufficiency or necessity.

PNS Theory in CoT Reasoning. Extending Pearl [46], we introduce the Probability of Necessary and Sufficient causes (PNS) framework to CoT reasoning, applying PN and PS at the step level rather than the model level as in Hüyük et al. [26]. This enables causal pruning of redundant steps, yielding minimal yet faithful CoTs. Relatedly, Yang et al. [72] formalize invariant learning through PNS estimation, offering a principled view of causal sufficiency and necessity in representation learning. Building on this foundation, recent work on LLM causal reasoning [29] extends similar principles to in-context learning. Our approach generalizes these insights to step-level reasoning dynamics, providing a model-agnostic and theoretically grounded alternative to heuristic compression methods.

Causal Necessity and Sufficiency in XAI. Prior studies have leveraged causal necessity and sufficiency to explain model behavior. LENS [60] identifies necessary and sufficient output conditions; Darwiche and Hirth [14] compute sufficient reasons via Decision-DNNF circuits; Mothilal et al. [41] generate diverse counterfactuals; Beckers [5] formalize sufficiency-based explanations for fairness; and Galhotra et al. [21] propose LEWIS, a probabilistic counterfactual method. Recent progress in causal representation learning [71, 70] further explores disentanglement and invariance through structural causal models and counterfactual reasoning, demonstrating how causal principles enable robust and interpretable representations. Building on these ideas, our PNS evaluation extends necessity and sufficiency analysis to LLM reasoning chains, using counterfactual rollouts to assess the causal faithfulness of CoT traces and mitigate overthinking.

3 Defining Causal Necessary and Sufficiency in CoT

3.1 Chain-of-Thought (CoT) Reasoning

Definition 1 (Chain-of-Thought (CoT) Reasoning [61]) Given an input $\mathbf{Q} = \mathbf{q}$, the Chain-of-Thought (CoT) reasoning process generates the final answer $\mathbf{A} = \mathbf{a}$ by sequentially deriving intermediate reasoning steps $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$. The probability of generating the answer given the question is defined as:

$$P(\mathbf{A} = \mathbf{a} \mid \mathbf{Q} = \mathbf{q}) \propto \int \underbrace{P(\mathbf{a} \mid \mathbf{s}_1, \dots, \mathbf{s}_n, \mathbf{q})}_{\text{Answer Generation}} \times \underbrace{\prod_{i=1}^n P(\mathbf{s}_i \mid \mathbf{s}_{<i}, \mathbf{q})}_{\text{CoT Generation}} d\mathbf{S}. \quad (1)$$

Explanation: $P(\mathbf{A} = \mathbf{a} \mid \mathbf{Q} = \mathbf{q})$ is the final answer probability. $P(\mathbf{a} \mid \mathbf{s}_1, \dots, \mathbf{s}_n, \mathbf{q})$ corresponds to the conditional probability of generating the final answer from the full reasoning trace. $\prod_{i=1}^n P(\mathbf{s}_i \mid \mathbf{s}_{<i}, \mathbf{q})$ models the sequential reasoning process. The integral marginalizes over all possible reasoning traces \mathbf{S} .

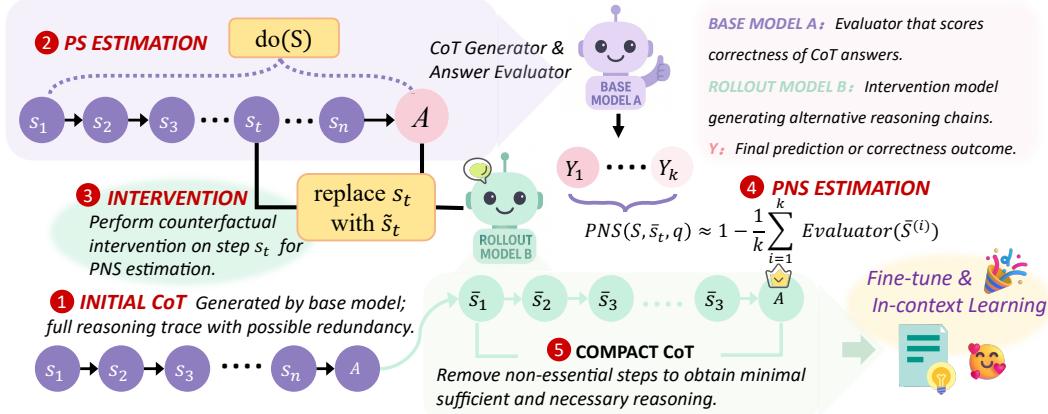


Figure 2: Causal Optimization Framework for CoT Reasoning. Our method identifies and retains only causally essential reasoning steps to form a compact CoT. (1) A base model generates the initial CoT trace, possibly containing redundant steps. (2) Sufficiency is estimated by checking if the full CoT leads to a correct answer. (3) For each step s_t , necessity is evaluated via counterfactual substitution \bar{s}_t using a rollout model, followed by answer scoring from the base model. (4) The Probability of Necessity and Sufficiency (PNS) is computed to measure causal contribution. (5) Non-essential steps are pruned to obtain a compact CoT, which is then used for fine-tuning or in-context learning.

3.2 Causal Necessary and Sufficiency in CoT

To rigorously characterize the causal significance of individual reasoning steps in Chain-of-Thought (CoT) reasoning, we propose formal definitions of causal sufficiency and causal necessity tailored to the structure and properties of CoT.

Definition 2 (Sufficiency, PS) *Sufficiency measures whether the reasoning chain $\mathbf{S} = (s_1, \dots, s_n)$ is sufficient to produce the correct answer y . Following the counterfactual definition in Pearl [46], the probability of sufficiency is defined as:*

$$PS(\mathbf{S}, q) = P(A_{do(\mathbf{S})} = y \mid A \neq y, \bar{\mathbf{S}}, q), \quad (2)$$

where $do(\mathbf{S})$ means the intervention which set the value of chain variable as \mathbf{S} , $A_{do(\mathbf{S})}$ denotes the counterfactual answer had the reasoning chain \mathbf{S} been used, and $A_{do(\bar{\mathbf{S}})}$ denotes the actual answer under the original reasoning $\bar{\mathbf{S}}$ (which could be null or incorrect). This captures the likelihood that inserting \mathbf{S} would have changed an incorrect answer to a correct one.

Definition 3 (Necessity, PN) *Necessity quantifies whether a specific reasoning step s_t is required for producing the correct answer $a = y$. Inspired by the counterfactual definition of necessity [46], we define the probability of necessity as:*

$$PN(\mathbf{S}, \bar{s}_t, q) = P(A_{do(s_{<t}, \bar{s}_t, s'>_t)} \neq y \mid A = y, \mathbf{S}, q), \quad (3)$$

where $s_{<t}$ denotes the set of all correct reasoning steps before position t , \bar{s}_t represents an incorrect or corrupted variant of the original step s_t . The counterfactual outcome $A_{do(\bar{s}_t)}$ is defined as the model's predicted answer when s_t is replaced by \bar{s}_t , and the subsequent steps $s'>_t$ are generated conditioned on this modified reasoning trajectory.

Definition 4 (Probability of Necessary and Sufficient Cause (PNS) in CoT) *Given a reasoning chain $\mathbf{S} = (s_1, \dots, s_n)$ that produces the correct answer $A = y$, and an alternative reasoning step \bar{s}_t at position t , let the counterfactual chain be defined as:*

$$\mathbf{S}' = (s_{<t}, \bar{s}_t, s'>_t),$$

where $s_{<t}$ denotes the preceding steps, and $s'>_t$ are subsequent steps possibly adapted to \bar{s}_t .

Inspired by the counterfactual definition [46], we define the Probability of Necessary and Sufficient Cause (PNS) for the step s_t as:

$$\text{PNS}(\mathbf{S}, \bar{\mathbf{s}}_t, \mathbf{q}) := P(\mathbf{A}_{\mathbf{S}} = \mathbf{y}, \mathbf{A}_{\mathbf{S}'} \neq \mathbf{y}). \quad (4)$$

This quantifies the probability that step \mathbf{s}_t is sufficient and necessary for the correct answer under a counterfactual. Identifiability results are in Appendix A.

4 Methodology - PNS Estimation for Improving Chain-of-Thought Reasoning

In practice, directly maximizing the full-chain Probability of Necessity and Sufficiency (PNS) for reconstructed CoTs is computationally intractable. To address this, we adopt a two-stage causal optimization strategy: we first enhance the chain-level Probability of Sufficiency (PS), followed by step-wise refinement via the node-level Probability of Necessity (PN). Guided by these criteria, we propose an iterative pruning framework—Algorithm 1—that removes and reorders steps to preserve only those that are both causally sufficient and necessary for producing the correct answer. The overall process is illustrated in Figure 2, which shows how initial CoTs are evaluated, intervened upon, and refined to produce minimal causal traces. These optimized CoTs are then used as high-quality exemplars for subsequent in-context learning and fine-tuning, enabling the base model to internalize which reasoning paths are truly essential.

4.1 PNS Estimation and Algorithm for Reconstructing CoT

Algorithm 1: Sufficient and Necessary Optimization of CoT

Input: Initial CoT chain \mathbf{S}_{init} , ground truth answer \mathbf{y} , query \mathbf{q} , threshold α
Output: Optimized CoT chain $\mathbf{S}_{\text{final}}$

```

 $\hat{\mathbf{y}}_{\text{init}} \leftarrow \text{Rollout}(\mathbf{S}_{\text{init}}, \mathbf{q});$ 
 $\text{PS} \leftarrow \mathbb{1}[\hat{\mathbf{y}}_{\text{init}} = \mathbf{y}];$ 
if  $\text{PS} = 1$  then
    Let  $\mathbf{S}_{\text{current}} \leftarrow \mathbf{S}_{\text{init}};$ 
    foreach step  $\mathbf{s}_t \in \mathbf{S}_{\text{current}}$  (processed in order) do
         $\bar{\mathbf{s}}_t \leftarrow \text{GenerateAlternative}(\mathbf{s}_{<t}^{\text{current}}, \mathbf{s}_t);$ 
         $V_{\text{scores}} \leftarrow \text{empty list};$ 
        for  $j \leftarrow 1 \dots k$  do
             $\bar{\mathbf{S}}^{(j)} \leftarrow \text{RolloutContinuation}(\mathbf{s}_{<t}^{\text{current}}, \bar{\mathbf{s}}_t, B); \quad // B \text{ is the rollout model};$ 
             $\text{Forms } (\mathbf{s}_{<t}^{\text{current}}, \bar{\mathbf{s}}_t, \mathbf{s}_{>t}^{(j)})$ 
            Ensure semantic disjointness of  $(\bar{\mathbf{s}}_t, \mathbf{s}_{>t}^{(j)})$  from original  $(\mathbf{s}_t, \mathbf{s}_{>t}^{\text{current}})$ ;
            Add  $V(\bar{\mathbf{S}}^{(j)})$  to  $V_{\text{scores}}; \quad // V \text{ is the validation model}$ 
             $\text{PNS}_{\text{val}}(\mathbf{s}_t) \leftarrow 1 - \frac{1}{k} \sum_{v \in V_{\text{scores}}} v;$ 
            if  $\text{PNS}_{\text{val}}(\mathbf{s}_t) > \alpha$  then
                Append  $\mathbf{s}_t$  to  $\mathbf{S}_{\text{final}}; \quad // \text{If } \mathbf{s}_t \text{ is deemed necessary, keep it}$ 
            else
                Skip  $\mathbf{s}_t; \quad // \text{Drop unnecessary step}$ 
        else
             $\mathbf{S}_{\text{final}} \leftarrow \mathbf{S}_{\text{init}}; \quad // \text{Original chain not sufficient}$ 
return  $\mathbf{S}_{\text{final}};$ 

```

Estimating PS (Chain-Level). We approximate CoT trace sufficiency as binary: $\text{PS} = 1$ if chain \mathbf{S} yields the correct answer (equivalent to $P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S}), \mathbf{q}) = 1$; Appendix A), else $\text{PS} = 0$. To improve PS, we repeatedly execute Algorithm 1 under the same question context. In each execution, the model samples an alternative CoT, and its PS is re-evaluated. This repeated sampling increases the likelihood of obtaining a CoT with higher PS. Lemma 1 (proof in Appendix A) establishes PNS identifiability (Definition 3.2) when $\text{PS} = 1$:

Lemma 1 (Identifiability of PNS under $P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S}), \mathbf{q}) = 1$) Assume:

1. *Perfect intervention with correct CoT $\mathbf{S} = (\mathbf{s}_{<t}, \mathbf{s}_t, \mathbf{s}_{>t})$ yields $P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S}), \mathbf{q}) = 1$.*
2. *Replacing step \mathbf{s}_t with incorrect $\bar{\mathbf{s}}_t$ and performing rollout $\mathbf{s}'_{>t}$ (from $\bar{\mathbf{s}}_t$) yields intervened chain $\bar{\mathbf{S}} = (\mathbf{s}_{<t}, \bar{\mathbf{s}}_t, \mathbf{s}'_{>t})$.*

Then, $PNS(\mathbf{S}, \bar{\mathbf{s}}_t, \mathbf{q}) = P(\mathbf{A}_{\text{do}(\mathbf{S})} = \mathbf{y}, \mathbf{A}_{\text{do}(\bar{\mathbf{S}})} \neq \mathbf{y} \mid \mathbf{q})$ simplifies to $1 - P(\mathbf{A} = \mathbf{y} \mid \text{do}(\bar{\mathbf{S}}), \mathbf{q})$, assuming perfect intervention and the nature of $\bar{\mathbf{S}}$ from Definition 3.2.

When $PS = 1$, PNS validity depends on $P(\mathbf{A} = \mathbf{y} \mid \text{do}(\bar{\mathbf{S}}), \mathbf{q})$, reflecting PN’s magnitude.

Estimating PN (Node-Level). If \mathbf{S} is sufficient ($PS(\mathbf{S}, \mathbf{q}) = 1$), Lemma 1 guides PN estimation for each node \mathbf{s}_t to evaluate $PNS(\mathbf{S}, \bar{\mathbf{s}}_t, \mathbf{q})$. We construct $\bar{\mathbf{s}}_t$ by removing \mathbf{s}_t ’s content and descendants. A rollout model B generates a revised downstream segment $(\bar{\mathbf{s}}_t, \mathbf{s}'_{>t})$, forming intervened chain $\bar{\mathbf{S}}^{(i)} = (\mathbf{s}_{<t}, \bar{\mathbf{s}}_t, \mathbf{s}'_{>t})$ for the i -th rollout, ensuring $\bar{\mathbf{s}}_t, \mathbf{s}'_{>t}$ are semantically disjoint from original components rooted at \mathbf{s}_t . Each $\bar{\mathbf{S}}^{(i)}$ is assessed by validation model V for coherence and logical integrity (beyond just final answer correctness). PNS is then computed via Monte-Carlo estimation over k rollouts:

$$PNS(\mathbf{S}, \bar{\mathbf{s}}_t, \mathbf{q}) \approx 1 - \frac{1}{k} \sum_{i=1}^k V(\bar{\mathbf{S}}^{(i)}). \quad (5)$$

Nodes with PNS score below threshold α (and their downstream nodes) are pruned iteratively until all retained nodes satisfy necessity.

Iterative Optimization. Algorithm 1 details the iterative optimization procedure. Starting from the initial CoT trace, we extract its chain and compute PNS. If the chain is sufficient, we perform necessity estimation for each node and prune the chain accordingly. The final optimized trace $\mathbf{S}_{\text{final}}$ consists only of reasoning steps that are both sufficient to produce the correct answer and necessary to preserve logical coherence.

Rollout Strategies for Intervention Chain $\bar{\mathbf{S}}^{(i)}$. We use three strategies (details/prompts in Appendix B) for generating semantically modified steps for $\bar{\mathbf{S}}^{(i)}$: (1) *Direct Rollout*: base model generates replacement from preceding context. (2) *Prompt-Based Rollout*: structured prompts guide base model substitutions. (3) *External Rollout*: a separate, stronger model generates replacements. Base/rollout models are consistent for (1)-(2); external rollouts use a more capable auxiliary.

4.2 PNS-Guided Reasoning Enhancement for In-Context Learning and Fine-Tuning

We use causally filtered CoT traces to improve LLMs under two paradigms: **In-Context Learning (ICL)** and **Supervised Fine-Tuning (SFT)**. In ICL, optimized CoTs are directly inserted into prompts to guide non-reasoning models. In SFT, we fine-tune reasoning-capable models using 1,229 high-quality CoTs. Results for both settings are reported in § 5.2.

5 Experiments

Our experiments are structured around two core questions:

RQ1: Can our method construct CoT datasets that enhance causal sufficiency and necessity? (§ 5.2.1)

RQ2: Can the causally optimized CoT data improve the performance of non-reasoning models via ICL, and further enhance reasoning-capable models through SFT? (§ 5.2.2)

5.1 Experimental Setup

Datasets. We evaluate on diverse reasoning benchmarks to ensure robustness across domains and difficulty levels. For mathematical reasoning, we use: (1) **GSM-8k** [10], with grade-school problems; (2) **MATH-500** [25], covering intermediate-level topics; and (3) **AIME**, with advanced competition problems up to 2025 [44, 11]. For commonsense reasoning, we use **CommonsenseQA** [53], a multiple-choice dataset requiring everyday inference.

Table 1: Experimental results for RQ1. Comparison of CoT reasoning performance before and after PNS-based optimization across QWEN and DEEPSEEK variants.

Method	GSM-8k			CommonsenseQA			MATH-500			AIME		
	Tokens (Initial/Final)	Steps (Initial/Final)	Acc. (Initial/Final)									
Qwen Variant (QwQ-32B-Preview & Qwen-2.5-72B-Instruct)												
Prompt-Based	113.8 →33.9	8.1 →2.3	90.0% →95.8%	109.2 →90.4	3.7 →3.0	69.7% →75.6%	281.8 →178.8	9.2 →5.5	82.6% →86.6%	531.3 →511.9	12.5 →12.3	16.7% →26.7%
Direct	113.8 →26.6	8.1 →2.0	90.0% →97.0%	109.2 →90.4	3.7 →3.0	69.7% →74.5%	281.8 →169.4	9.2 →5.1	82.6% →87.4%	531.3 →522.7	12.5 →12.3	16.7% →23.3%
External	356.4 →58.9	23.9 →3.0	93.3% →97.9%	474.2 →215.9	17.8 →7.4	83.2% →88.0%	743.2 →200.7	50.3 →11.0	87.6% →94.3%	1719.4 →1479.4	108.7 →76.7	43.3% →56.7%
DeepSeek Variant (DeepSeek-R1 & DeepSeek-V3)												
Prompt-Based	137.3 →29.6	5.4 →1.4	95.0% →97.3%	191.0 →70.9	6.2 →2.7	83.0% →85.3%	387.6 →163.4	16.2 →6.2	85.9% →92.0%	2438.8 →2195.4	120.3 →100.9	25.0% →25.0%
Direct	137.3 →29.2	5.4 →1.3	95.0% →97.0%	191.0 →69.8	6.2 →2.5	83.0% →85.3%	387.6 →161.2	16.2 →5.9	85.9% →91.6%	2438.8 →2082.6	120.3 →96.9	25.0% →25.0%
External	451.6 →135.1	6.7 →3.2	99.0% →99.9%	368.2 →167.9	7.6 →3.1	87.1% →94.9%	828.8 →214.5	29.7 →6.1	93.2% →97.6%	3052.5 →1639.3	157.8 →68.3	79.2% →92.6%

Evaluation Metrics. We assess (i) reasoning efficiency—measured by token length and step length—and (ii) accuracy, comprising PS(chain) and final-answer accuracy. Token length is computed via space-delimited tokenization; step length counts steps separated by a double newline for consistency. Accuracy is the average fraction of correctly answered instances on the test set, i.e., Accuracy = $\frac{\# \text{ correctly answered instances}}{\# \text{ instances in the test set}}$.

Baselines. For RQ1, we compare unoptimized CoT traces from the base model with PNS-optimized versions (via § 4.1) in terms of reasoning efficiency, accuracy, and average PNS. A representative example is shown in Figure 3, with full results in Appendix G.

For RQ2, ICL baselines include: (1) **Standard**, few-shot with original (often redundant) CoTs; (2) **Fast-Solve**, concise yet complete reasoning; (3) **Reduction** [19], shortcut conclusions; (4) **CoD** [67], minimal key phrases; and (5) **Ours-ICL**, few-shot with PNS-optimized, causally essential steps. Details and prompts are in Appendix C. For SFT, we compare: **Original**, the base model; **Noncausal**, fine-tuned on raw CoTs; and **Causal**, fine-tuned on the same CoTs after PNS-based pruning.

Models. For RQ1, we use **Qwen-2.5-72B-Instruct** [55] as both the base and rollout model. In the **External** variant, **QwQ-32B-Preview** [56] is used as base, with rollout unchanged. For DeepSeek, DeepSeek-V3 [16] serves as both base and rollout in standard settings, while the external variant uses DeepSeek-R1 [17] as base and V3 as rollout. PNS evaluations share the same configuration.

RQ2 ICL experiments primarily use **Qwen-2.5-72B-Instruct**, with additional results from **Qwen-2.5-7B-Instruct** [55], **Llama-3.1-8B-Instruct** [1], and **DeepSeek-V3**. SFT experiments fine-tune **DeepSeek-R1-Distill-Qwen-1.5B**, **DeepScaleR-1.5B-Preview** [37], and **Phi-4-mini-reasoning** [66]. Training details are in Appendix D.

5.2 Main Results

5.2.1 PNS Optimization on CoT Trajectories

We apply our method to CoT traces from Qwen and DeepSeek variants. As shown in Table 1, our PNS-based algorithm reduces both token and step lengths while improving accuracy⁴, indicating effective removal of redundant reasoning.

We compare average PNS values before and after optimization across tasks (AIME and CommonsenseQA) and models (**Qwen-2.5-72B-Instruct**, **DeepSeek-R1**). As shown in Figure 3, PNS values consistently increase after optimization, confirming that the retained steps are more causally sufficient and necessary. The figure illustrates results on 15 sampled questions per setting; more comprehensive results across larger test sets are provided in Appendix G.

⁴Inference for reasoning models was performed using VLLM. The `max-tokens` is 16,384.

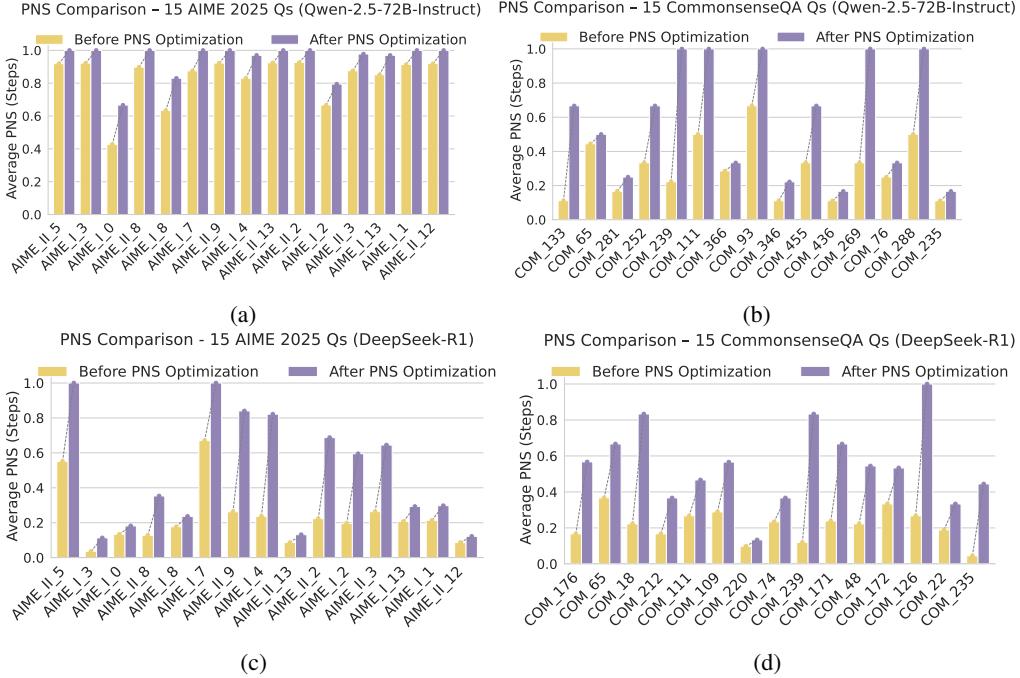


Figure 3: Average PNS values before and after optimization across different models and datasets. Each subfigure displays PNS improvements across 15 sampled problems: (a) Qwen-2.5-72B-Instruct evaluated on AIME, (b) Qwen-2.5-72B-Instruct on CommonsenseQA, (c) DeepSeek-R1 evaluated on AIME, and (d) DeepSeek-R1 on CommonsenseQA. PNS-optimization CoTs exhibit consistently higher PNS values, indicating an increased necessity for retained steps.

We conducted a human quality evaluation of 50 CoTs: 84% were judged both sufficient and necessary (S&N), and only 6% insufficient (NbI); see Appendix I for details.

These findings indicate that our optimized CoTs are not only more concise and accurate, but also demonstrate enhanced causal sufficiency and necessity. Notably, the average PNS per step increases after optimization, suggesting that the retained reasoning steps are more integral to producing correct answers—each step contributes more critically to the final outcome than before.

5.2.2 Enhancing LLMs via ICL and SFT with Optimized CoT

We investigate enhancing LLM performance with optimized CoT data via in-context learning (ICL) and supervised fine-tuning (SFT).

In-Context Learning with Optimized CoT. Using optimized CoT traces for ICL with non-fine-tuned LLMs (Table 2), Ours-ICL balances reasoning efficiency and accuracy. Compared to Standard CoT, it consistently reduces token/step usage (often >50%) with minimal/no accuracy loss. For instance, on GSM-8k with DeepSeek-V3, Ours-ICL improves accuracy (97.6% to 99.9%) while cutting tokens by 67%; with Llama-3.1-8B-Instruct on MATH-500, accuracy increases by 7.1 points (to 54.8%) with more concise reasoning. Ours-ICL also surpasses baselines like Fast-Solve and Reduction [19] in accuracy with comparable or better efficiency. On GSM-8k with Qwen-2.5-72B-Instruct, our method (99.5%, 65.3 tokens) outperforms Fast-Solve (91.7%, 72.8 tokens) and Reduction (84.7%, 114.1 tokens). Unlike aggressive pruning (e.g., CoD [67]), Ours-ICL maintains substantially higher accuracy, especially on complex tasks (e.g., MATH-500: 96.2% vs. 55.6% with DeepSeek-V3).

Limitations for ICL. Despite its effectiveness with non-reasoning models, ICL using optimized CoT is sensitive to prompt/example selection. Its benefits diminish on complex tasks (e.g., MATH-500), and performance is constrained by the fixed parameters of such models. In contrast, SFT is more impactful for reasoning-capable models, allowing deeper integration of reasoning patterns.

Table 2: Experimental results for RQ2 (ICL on Non-Reasoning Models). Lower is better (\downarrow) for Tokens/Steps, higher is better (\uparrow) for Acc. Change rates (%) relative to the average of methods for that metric/dataset/model are in parentheses. Cells significantly better than average are colored blue (deeper for greater improvement).

Method	CommonsenseQA			GSM-8k			MATH-500		
	Tokens \downarrow	Steps \downarrow	Acc. \uparrow	Tokens \downarrow	Steps \downarrow	Acc. \uparrow	Tokens \downarrow	Steps \downarrow	Acc. \uparrow
DeepSeek-V3									
Standard	177.5 (+90.9%)	5.7 (+54.1%)	83.8% (+1.5%)	157.3 (+83.3%)	7.4 (+51.0%)	97.6% (+3.0%)	598.6 (+90.4%)	26.7 (+88.0%)	93.2% (+9.9%)
Fast-Solve	120.1 (+29.1%)	4.7 (+27.0%)	82.0% (-0.7%)	86.3 (+0.6%)	4.9 (0.0%)	95.1% (+0.3%)	329.2 (+4.7%)	13.6 (-4.2%)	87.2% (+2.8%)
Reduction [19]	103.4 (+11.2%)	3.6 (-2.7%)	83.1% (+0.6%)	104.4 (+21.7%)	6.0 (+22.4%)	97.3% (+2.6%)	476.3 (+51.5%)	22.0 (+54.9%)	91.6% (+8.0%)
CoD [67]	19.3 (-79.2%)	2.0 (-45.9%)	80.7% (-3.3%)	28.7 (-66.5%)	1.8 (-63.3%)	84.0% (-11.4%)	31.2 (-90.1%)	2.1 (-85.2%)	55.6% (-34.4%)
Ours-ICL	44.7 (-51.9%)	2.7 (-27.0%)	83.6% (+1.2%)	52.2 (-39.2%)	4.3 (-12.2%)	99.9% (+5.4%)	136.7 (-56.5%)	6.4 (-54.9%)	96.2% (+13.4%)
Qwen-2.5-72B-Instruct									
Standard CoT	109.1 (+43.0%)	3.7 (+27.6%)	78.2% (-1.1%)	113.8 (+47.8%)	8.0 (+81.8%)	93.6% (+4.6%)	281.8 (+51.9%)	9.2 (+27.8%)	84.0% (+16.8%)
Fast-Solve	59.7 (-21.7%)	2.0 (-31.0%)	67.0% (-15.3%)	72.8 (-5.5%)	2.7 (-38.6%)	91.7% (+2.5%)	192.4 (+3.7%)	7.3 (+1.4%)	69.8% (-2.9%)
Reduction [19]	116.6 (+52.8%)	3.5 (+20.7%)	84.9% (+7.3%)	114.1 (+48.2%)	4.8 (-9.1%)	84.7% (-5.4%)	233.3 (+25.8%)	9.4 (+30.6%)	72.4% (+0.7%)
CoD [67]	14.4 (-81.1%)	2.1 (-27.6%)	82.3% (+4.0%)	18.8 (-75.6%)	1.1 (-74.9%)	78.1% (-12.7%)	23.0 (-87.6%)	1.2 (-83.3%)	52.0% (-27.6%)
Ours-ICL	81.6 (+7.0%)	3.4 (+17.2%)	83.0% (+4.9%)	65.3 (-15.2%)	5.3 (+20.5%)	99.5% (+11.2%)	196.9 (+6.1%)	8.9 (+23.6%)	81.2% (+13.0%)
Qwen-2.5-7B-Instruct									
Standard CoT	209.8 (+68.0%)	7.2 (+46.9%)	70.4% (-6.5%)	149.7 (+48.8%)	7.2 (+33.3%)	85.1% (+0.6%)	263.5 (+43.0%)	9.7 (+19.8%)	71.0% (+12.9%)
Fast-Solve	120.1 (-3.8%)	5.0 (+2.0%)	76.5% (+1.6%)	108.4 (+7.8%)	5.5 (+1.9%)	83.6% (-1.2%)	200.6 (+8.8%)	8.4 (+3.7%)	64.6% (+2.7%)
Reduction [19]	178.5 (+42.9%)	6.8 (+38.8%)	74.4% (-1.2%)	131.1 (+30.3%)	6.4 (+18.5%)	84.3% (-0.4%)	232.2 (+26.0%)	9.1 (+12.3%)	71.8% (+14.1%)
CoD [67]	16.8 (-86.5%)	1.9 (-60.7%)	77.4% (+2.8%)	30.3 (-69.9%)	2.9 (-46.3%)	75.7% (-10.5%)	50.3 (-72.7%)	5.8 (-28.3%)	34.6% (-45.0%)
Ours-ICL	99.1 (-20.7%)	3.8 (-22.4%)	77.6% (+3.1%)	83.4 (-17.1%)	4.8 (-11.1%)	94.1% (+11.2%)	174.7 (-5.2%)	7.7 (-4.9%)	72.6% (+15.4%)
Llama-3.1-8B-Instruct									
Standard CoT	169.3 (+33.5%)	7.3 (+15.9%)	72.2% (+2.8%)	182.8 (+29.2%)	7.9 (-1.2%)	79.2% (-0.1%)	741.3 (+55.2%)	46.0 (+39.8%)	47.6% (+6.7%)
Fast-Solve	140.7 (+10.9%)	7.9 (+25.4%)	69.0% (-0.7%)	170.6 (+20.6%)	9.5 (+18.2%)	72.0% (-9.2%)	453.1 (-5.1%)	30.9 (-6.1%)	46.4% (+4.0%)
Reduction [19]	143.2 (+13.0%)	5.8 (-7.9%)	69.9% (-0.4%)	129.1 (-8.8%)	6.2 (-22.4%)	82.8% (+4.4%)	515.6 (+8.0%)	34.0 (+3.3%)	46.2% (+3.6%)
CoD [67]	60.6 (-52.2%)	3.5 (-44.3%)	67.6% (-3.7%)	96.7 (-31.6%)	8.5 (+6.2%)	69.2% (-12.7%)	312.5 (-34.6%)	29.2 (-11.2%)	28.0% (-37.2%)
Ours-ICL	120.3 (-5.1%)	7.1 (+12.7%)	72.1% (+2.7%)	128.3 (-9.3%)	8.1 (+1.2%)	93.1% (+17.4%)	365.9 (-23.4%)	24.2 (-26.4%)	54.8% (+22.9%)

Supervised Fine-Tuning with Optimized CoT. We fine-tune reasoning models on 1,229 PNS-selected CoT traces from MATH [25], MMLU [24], ZebraLogicBench [34], CommonsenseQA [53], and AIME (pre-2024) [44]. All traces were manually checked for causal sufficiency, necessity, concision, and quality; CoTs selected by our method were not edited—only verified. Table 3 demonstrates Causal-CoT’s consistent outperformance over baselines. On CommonsenseQA, it improves accuracy to 47.2% (from 37.6% on DeepSeek-R1-Distill-Qwen-1.5B, and from 41.3% on DeepScaleR-1.5B-Preview) while halving reasoning steps. On MATH-500, it achieves 78.2% accuracy (33.1 steps) versus the Original model’s 76.4% (77.2 steps). Even on difficult tasks like AIME25 (low absolute performance⁵), Causal-CoT significantly cuts reasoning length (e.g., DeepSeek-R1-Distill-Qwen-1.5B: 212.4 to 95.4 tokens). The Causal variant also matches/exceeds Noncausal fine-tuning with substantially fewer steps/tokens. For GSM-8k, it reaches 86.2% accuracy (11.6 steps), while Noncausal needs 15 steps for 86.1% (DeepScaleR-1.5B-Preview).

Implications of SFT Results. SFT on PNS-selected traces yields consistent gains in accuracy and reasoning efficiency, even with small Causal-CoT datasets, confirming the high supervision value of enforcing causal sufficiency and necessity. The token overhead from PNS filtering is a one-time curation cost; after fine-tuning, inference becomes cheaper because the model generates concise, accurate CoTs without stepwise rollouts or post-hoc pruning. The strength of our approach lies in reshaping the training distribution toward high-PNS evidence, enabling the model to internalize causally meaningful, non-redundant reasoning patterns. At test time, this manifests as streamlined, interpretable CoTs that improve both efficiency and reliability.

6 Conclusion

This work successfully incorporates PNS into CoT reasoning. The developed method systematically prunes unnecessary reasoning steps, leading to significant improvements in reasoning efficiency, while maintaining or even enhancing the accuracy of the outcomes. The effectiveness of the method has been verified across both in-context learning and supervised fine-tuning scenarios.

Limitation and Future Work. Limitations include potential performance decreases on highly complex tasks. Key challenges involve selecting optimal pruning thresholds, managing PNS esti-

⁵Inference performed via VLLM; The max-tokens is 16,384.

Table 3: Experimental results for RQ2 (SFT on Reasoning Models). For Noncausal and Causal methods, change rates (%) are reported relative to the “Original” for each model/dataset. Cells with notable improvements over the original are highlighted.

Method	CommonsenseQA			GSM-8k			MATH-500			AIME25		
	Tokens↓	Steps↓	Acc.↑	Tokens↓	Steps↓	Acc.↑	Tokens↓	Steps↓	Acc.↑	Tokens↓	Steps↓	Acc.↑
DeepSeek-R1-Distill-Qwen-1.5B												
Original	751.3 (baseline)	21.4 (baseline)	37.6% (baseline)	332.1 (baseline)	14.4 (baseline)	77.9% (baseline)	1441.8 (baseline)	77.2 (baseline)	76.4% (baseline)	4002.3 (baseline)	212.4 (baseline)	23.3% (baseline)
Noncausal	1271.3 (+69.2%)	27.9 (+30.4%)	43.2% (+14.9%)	621.4 (+87.1%)	16.5 (+14.6%)	83.4% (+7.1%)	1456.6 (+1.0%)	54.7 (-29.1%)	81.6% (+6.8%)	3796.3 (-5.1%)	116.4 (-45.2%)	20.0% (-14.2%)
Causal (Ours)	740.0 (-1.5%)	10.3 (-51.9%)	47.2% (+25.5%)	327.8 (-1.3%)	12.8 (-11.1%)	84.2% (+8.1%)	911.9 (-36.7%)	33.1 (-57.1%)	78.2% (+2.4%)	2948.3 (-26.3%)	95.4 (-55.1%)	23.3% (0.0%)
DeepScaleR-1.5B-Preview												
Original	646.7 (baseline)	22.1 (baseline)	41.3% (baseline)	716.3 (baseline)	27.5 (baseline)	87.5% (baseline)	1325.0 (baseline)	77.4 (baseline)	84.6% (baseline)	4897.2 (baseline)	356.0 (baseline)	20.0% (baseline)
Noncausal	611.5 (-5.4%)	21.7 (-1.8%)	40.8% (-1.2%)	604.8 (-15.6%)	15.0 (-45.5%)	86.1% (-1.6%)	1445.5 (+9.1%)	52.4 (-32.3%)	82.8% (-2.1%)	5709.8 (+16.6%)	243.2 (-31.7%)	16.7% (-16.5%)
Causal (Ours)	601.2 (-7.0%)	11.2 (-49.3%)	47.2% (+14.3%)	394.6 (-44.9%)	11.6 (-57.8%)	86.2% (-1.5%)	1041.4 (-21.4%)	36.3 (-53.1%)	91.7% (+8.4%)	2015.4 (-58.8%)	53.6 (-85.0%)	20.0% (0.0%)
Phi-4-mini-reasoning												
Original	935.7 (baseline)	18.6 (baseline)	72.4% (baseline)	783.3 (baseline)	18.6 (baseline)	92.6% (baseline)	1743.5 (baseline)	62.1 (baseline)	85.8% (baseline)	6544.0 (baseline)	274.3 (baseline)	30.0% (baseline)
Noncausal	949.6 (+1.5%)	14.9 (-19.9%)	66.4% (-8.3%)	566.9 (-27.6%)	14.3 (-23.1%)	89.8% (-3.0%)	2042.8 (+17.2%)	65.8 (+5.9%)	59.8% (-30.3%)	8297.8 (+26.8%)	359.3 (+31.0%)	23.3% (-22.3%)
Causal (Ours)	920.3 (-1.6%)	14.1 (-24.2%)	72.9% (+0.7%)	517.2 (-34.0%)	14.0 (-24.7%)	92.4% (-0.2%)	1031.0 (-40.9%)	28.7 (-53.8%)	86.7% (+1.0%)	4140.0 (-36.7%)	112.0 (-59.2%)	30.0% (0.0%)

mation costs, and ensuring counterfactual generation quality. Future work will aim to address these limitations, focusing on improving the causal fidelity and overall performance of LLM reasoning.

Acknowledgments

X. Yu was supported in part by the National Natural Science Foundation of China (Nos. 62472306, 62441221, and 62206116), Tianjin University’s 2024 Special Project on Disciplinary Development (No. XKJS-2024-5-9), the Tianjin University Talent Innovation Reward Program for Literature & Science Graduate Students (No. C1-2022-010), and the Henan Province Key Research and Development Program (No. 251111210500). H. Li was supported in part by National Natural Science Foundation of China (No. 623B2002).

Author Contributions

This work was completed through the joint efforts of all authors. X.Y and Z.W were responsible for the main experiment design, model implementation, and manuscript writing. M.Y contribute to idea formulate, experiment design, theoretical analysis and manuscript writing. L.Y contribute to experimental design. H.L contribute to idea formulate and manuscript writing. The other authors contributed to regular discussion.

References

- [1] AI@Meta. Llama 3 model card, 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- [2] Ameen Ali, Lior Wolf, and Ivan Titov. Mitigating copy bias in in-context learning through neuron pruning. *arXiv preprint arXiv:2410.01288*, 2024.
- [3] Dean Allemang and Juan Sequeda. Increasing the llm accuracy for question answering: Ontologies to the rescue! *arXiv preprint arXiv:2405.11706*, 2024.

- [4] Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Dushyant Singh Sengar, Mayank Jindal, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. Cause and effect: can large language models truly understand causality? In *Proceedings of the AAAI Symposium Series*, volume 4, pages 2–9, 2024.
- [5] Sander Beckers. Causal explanations and xai. In *Conference on causal learning and reasoning*, pages 90–109. PMLR, 2022.
- [6] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczek, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690, 2024.
- [7] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for $2+3=?$ on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.
- [8] Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert Van Rooij, Kun Zhang, and Zhouchen Lin. Empowering llms with logical reasoning: A comprehensive survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025.
- [9] Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024.
- [10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reichihiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [11] OpenCompass Contributors. Aime 2025 dataset. <https://huggingface.co/datasets/opencompass/AIME2025>, 2025. Accessed: 2025-05-14.
- [12] Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, et al. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*, 2025.
- [13] Yingqian Cui, Pengfei He, Jingying Zeng, Hui Liu, Xianfeng Tang, Zhenwei Dai, Yan Han, Chen Luo, Jing Huang, Zhen Li, et al. Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2502.13260*, 2025.
- [14] Adnan Darwiche and Auguste Hirth. On the (complete) reasons behind decisions. *Journal of Logic, Language and Information*, 32(1):63–88, 2023.
- [15] C Nicolò De Sabbata, Theodore R Sumers, and Thomas L Griffiths. Rational metareasoning for large language models. *arXiv preprint arXiv:2410.05563*, 2024.
- [16] DeepSeek-AI. Deepseek-v3 technical report, 2024. URL <https://arxiv.org/abs/2412.19437>.
- [17] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- [18] Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*, 2023.
- [19] Mengru Ding, Hanmeng Liu, Zhizhang Fu, Jian Song, Wenbo Xie, and Yue Zhang. Break the chain: Large language models can be shortcut reasoners. *arXiv preprint arXiv:2406.06580*, 2024.
- [20] Xidong Feng, Yicheng Luo, Ziyan Wang, Hongrui Tang, Mengyue Yang, Kun Shao, David Mguni, Yali Du, and Jun Wang. Chessgpt: Bridging policy learning and language modeling. *Advances in Neural Information Processing Systems*, 36:7216–7262, 2023.

- [21] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, pages 577–590, 2021.
- [22] Tingxu Han, Chunrong Fang, Shiyu Zhao, Shiqing Ma, Zhenyu Chen, and Zhenting Wang. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*, 2024.
- [23] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.
- [24] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [25] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- [26] Alihan Hüyük, Xinnuo Xu, Jacqueline Maasch, Aditya V Nori, and Javier González. Reasoning elicitation in language models via counterfactual feedback. *arXiv preprint arXiv:2410.03767*, 2024.
- [27] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [28] Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, et al. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. *arXiv preprint arXiv:2404.07103*, 2024.
- [29] Jiarui Jin, Yuwei Wu, Haoxuan Li, Xiaoting He, Weinan Zhang, Yiming Yang, Yong Yu, Jun Wang, and Mengyue Yang. Large language models are demonstration pre-selectors for themselves. In *Forty-second International Conference on Machine Learning*, 2025.
- [30] Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*, 2024.
- [31] Ziqi Jin and Wei Lu. Self-harmonized chain of thought. *arXiv preprint arXiv:2409.04057*, 2024.
- [32] Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. C3ot: Generating shorter chain-of-thought without compromising effectiveness. *arXiv preprint arXiv:2412.11664*, 2024.
- [33] Leo Li, Ye Luo, and Tingyou Pan. Openai-o1 ab testing: Does the o1 model really do good reasoning in math problem solving? *arXiv preprint arXiv:2411.06198*, 2024.
- [34] Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025.
- [35] Yule Liu, Jingyi Zheng, Zhen Sun, Zifan Peng, Wenhan Dong, Zeyang Sha, Shiwen Cui, Weiqiang Wang, and Xinlei He. Thought manipulation: External thought can be efficient for large reasoning models. *arXiv preprint arXiv:2504.13626*, 2025.
- [36] Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*, 2025.
- [37] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl, 2025. Notion Blog.

- [38] Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*, 2025.
- [39] Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. Cot-valve: Length-compressible chain-of-thought tuning. *arXiv preprint arXiv:2502.09601*, 2025.
- [40] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.
- [41] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.
- [42] Tergel Munkhbat, Namgyu Ho, Seohyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. Self-training elicits concise reasoning in large language models. *arXiv preprint arXiv:2502.20122*, 2025.
- [43] OpenAI. Learning to reason with llms, 2025. URL <https://openai.com/index/learning-to-reason-with-llms/>.
- [44] OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, et al. Openai o1 system card, 2024. URL <https://arxiv.org/abs/2412.16720>.
- [45] Soumyasundar Pal, Didier Chételat, Yingxue Zhang, and Mark Coates. Hint marginalization for improved reasoning in large language models. *arXiv preprint arXiv:2412.13292*, 2024.
- [46] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [47] Haritz Puerto, Tilek Chubakov, Xiaodan Zhu, Harish Tayyar Madabushi, and Iryna Gurevych. Fine-tuning with divergent chains of thought boosts reasoning through self-correction in language models. *OpenReview*, 2024.
- [48] Team Qwen. Qwq: Reflect deeply on the boundaries of the unknown. <https://qwenlm.github.io/blog/qwq-32b-preview/>, 2024. Accessed: March 2025.
- [49] Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. Trust me, i'm wrong: High-certainty hallucinations in llms. *arXiv preprint arXiv:2502.12964*, 2025.
- [50] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*, 2024.
- [51] Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. *Advances in Neural Information Processing Systems*, 37:125019–125049, 2024.
- [52] Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. Fast best-of-n decoding via speculative rejection. *arXiv preprint arXiv:2410.20290*, 2024.
- [53] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL <https://arxiv.org/abs/1811.00937>.
- [54] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.
- [55] Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL <https://qwenlm.github.io/blog/qwen2.5/>.

- [56] Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, November 2024. URL <https://qwenlm.github.io/blog/qwq-32b-preview/>.
- [57] Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory. *arXiv preprint arXiv:2411.11984*, 2024.
- [58] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.
- [59] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- [60] David S Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. Local explanations via necessity and sufficiency: Unifying theory and practice. In *Uncertainty in Artificial Intelligence*, pages 1382–1392. PMLR, 2021.
- [61] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [62] Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, et al. A comparative study on reasoning patterns of openai's o1 model. *arXiv preprint arXiv:2410.13639*, 2024.
- [63] Tong Wu, Chong Xiang, Jiachen T Wang, and Prateek Mittal. Effectively controlling reasoning models through thinking intervention. *arXiv preprint arXiv:2503.24370*, 2025.
- [64] Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, et al. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought. *arXiv preprint arXiv:2501.04682*, 2025.
- [65] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.
- [66] Haoran Xu, Baolin Peng, Hany Awadalla, Dongdong Chen, Yen-Chun Chen, Mei Gao, Young Jin Kim, Yunsheng Li, Liliang Ren, Yelong Shen, Shuhang Wang, Weijian Xu, Jianfeng Gao, and Weizhu Chen. Phi-4-mini-reasoning: Exploring the limits of small reasoning language models in math, 2025. URL <https://arxiv.org/abs/2504.21233>.
- [67] Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*, 2025.
- [68] Xiao Xue, Xiangning Yu, Deyu Zhou, Xiao Wang, Chongke Bi, Shufang Wang, and Fei-Yue Wang. Computational experiments for complex social systems: Integrated design of experiment system. *IEEE/CAA Journal of Automatica Sinica*, 11(5):1175–1189, 2024.
- [69] Hao Yang, Qianghua Zhao, and Lei Li. Chain-of-thought in large language models: Decoding, projection, and activation. *arXiv preprint arXiv:2412.03944*, 2024.
- [70] Mengyue Yang, Quanyu Dai, Zhenhua Dong, Xu Chen, Xiuqiang He, and Jun Wang. Top-n recommendation with counterfactual user preference simulation. *CIKM '21*, page 2342–2351, 2021. doi: 10.1145/3459637.3482305.
- [71] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021.
- [72] Mengyue Yang, Zhen Fang, Yonggang Zhang, Yali Du, Furui Liu, Jean-Francois Ton, Jianhong Wang, and Jun Wang. Invariant learning via probability of sufficient and necessary causes. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, pages 79832–79857, 2023.

- [73] Wen Yang, Kai Fan, and Minpeng Liao. Markov chain of thought for efficient mathematical reasoning. *arXiv preprint arXiv:2410.17635*, 2024.
- [74] Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*, 2025.
- [75] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.
- [76] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [77] Xiangning Yu, Xiao Xue, Deyu Zhou, Li Fang, and Zhiyong Feng. Beyond traditional metrics: The power of value entropy in multidimensional evaluation of the service ecosystem. In *2024 IEEE International Conference on Web Services (ICWS)*, pages 611–621. IEEE, 2024.
- [78] Murong Yue, Wenlin Yao, Haitao Mi, Dian Yu, Ziyu Yao, and Dong Yu. Dots: Learning to reason dynamically in llms via optimal reasoning trajectories search. *arXiv preprint arXiv:2410.03864*, 2024.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer **[Yes]** , **[No]** , or **[NA]** .
- **[NA]** means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "**[Yes]**" is generally preferable to "**[No]**", it is perfectly acceptable to answer "**[No]**" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "**[No]**" or "**[NA]**" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer **[Yes]** to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”.**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claims presented in the abstract and introduction accurately reflect the scope and contributions of our work. Specifically, we claim that our proposed PNS-based CoT optimization improves reasoning efficiency and accuracy, which is supported by both theoretical formulation and empirical results discussed in § 4 and § 5.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We explicitly discuss limitations in § 6, noting reduced effectiveness on high-complexity tasks like AIME. This is attributed to the use of a fixed pruning threshold α , which may be suboptimal across diverse reasoning domains.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide a formal definition of the Probability of Necessity and Sufficiency (PNS) and clearly state the assumptions required for its estimation within the CoT framework. Key theoretical formulations and pruning criteria are detailed in § 4, with all assumptions explicitly stated. Formal justifications and derivations are included in the appendix to support our main results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All experimental settings—including datasets, model variants, training configurations, evaluation metrics, and optimization parameters—are described in detail in § 5 and Appendix D. A public GitHub repository including code and processed data will be released upon acceptance to further support reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use publicly available datasets and will release code and processed data upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide full experimental details in § 5 and the appendix, including dataset splits, prompt formatting, model variants, inference settings (e.g., max tokens, temperature).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We clearly specify our datasets and rollout methodology; all parameters are publicly disclosed, and results are obtained through repeated testing across multiple datasets.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide detailed descriptions of the compute resources in Appendix ??.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our research adheres fully to the NeurIPS Code of Ethics. All datasets used are publicly available, no private or sensitive data is involved, and we have considered potential social impacts in our discussion.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer:[NA]

Justification: This work is fundamental research on reasoning optimization and does not directly involve applications with immediate societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work does not involve the release of new pretrained models or scraped datasets that pose a high risk of misuse. We evaluate existing public models and operate on curated, publicly available datasets with no sensitive content.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets used in this work, including datasets, code, and models, are publicly available. We have properly credited the original creators and ensured that all licenses and terms of use are respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release accompanying code. All assets are documented with details on format, usage, and limitations, and are provided via an anonymized GitHub repository during review, with public release planned upon publication.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs are used in this paper solely for text polishing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Appendix

Appendix Table of Contents

- A. Theoretical Analysis
 - B. Intervention Prompts
 - C. In-Context Learning Prompts
 - D. Supervised Fine-Tuning (SFT) Hyperparameter Settings
 - E. In-Context Learning: Case study
 - F. Full Version of Related Work
 - G. Additional PNS Comparison Results
 - H. Computational Complexity Analysis
 - I. Qualitative Analysis of Whether the Final Reasoning Is Sufficient and Necessary
 - J. Validator Accuracy and Robustness
 - K. Supplementary Experimental Results
-

A Theoretical Analysis

A.1 Exogeneity and Monotonicity Conditions

For the calculation of the Probability of Necessary and Sufficient (PNS) conditions, it is essential that the exogeneity and monotonicity conditions are satisfied.

Definition 5 (CoT Exogeneity) *When generating each reasoning step s_t , the generation of the current step depends only on the previous reasoning steps $s_{<t}$ and the question q , and not on any external variables. Formally:*

$$P(s_t | \text{do}(s_{<t}), q) = P(s_t | s_{<t}, q) \quad (\text{A.1})$$

Where s_t is the current reasoning step, $s_{<t}$ denotes all previous reasoning steps, and q is the input question. According to the definition of exogeneity, the intervention probability can be evaluated by the conditional probability.

Assumption 1 (Monotonicity for CoT) *Inspired by the monotonicity assumption in causal inference [46], we define monotonicity for Chain-of-Thought (CoT) reasoning as follows. Let $S = (s_1, \dots, s_n)$ be a reasoning chain leading to the correct answer $A = y$, and let $\bar{S} = (s_1, \dots, s_{t-1}, \bar{s}_t, s'_{t+1}, \dots, s'_n)$ denote a modified reasoning chain where only step t is altered to an incorrect step \bar{s}_t and subsequent steps $s'_{>t}$ are rolled out. CoT reasoning satisfies monotonicity if and only if, for every reasoning step t and for every possible alteration \bar{s}_t , the following holds:*

$$P(A_S \neq y, A_{\bar{S}} = y) = 0.$$

This condition states that altering a reasoning step to an incorrect version cannot result in correcting a previously incorrect final answer. Equivalently, whenever the modified chain \bar{S} yields a correct answer, the original unaltered chain S must also yield a correct answer, ensuring monotonic progression toward correctness.

A.2 Identifiability of PNS in CoT Under Monotonicity Assumption

Lemma 2 (Identifiability of PNS under downstream-adaptive reasoning) *Assume that the Chain-of-Thought (CoT) reasoning process satisfies both the Exogeneity (Definition A.1) and Monotonicity (Assumption 1) assumptions. Let the correct reasoning chain be denoted by $S = (s_{<t}, s_t, s_{>t})$, and let an alternative corrupted step \bar{s}_t induce a modified future reasoning sequence $s'_{>t}$, resulting in the altered chain $S' = (s_{<t}, \bar{s}_t, s'_{>t})$. Then the Probability of Necessary and Sufficient Cause (PNS) for the reasoning chain is identifiable and satisfies:*

$$\text{PNS}(S, \bar{s}_t) = P(A_S = y) - P(A_{S'} = y) = P(A = y | \text{do}(S)) - P(A = y | \text{do}(S')).$$

Note: $\text{do}(S')$ is equivalent to $\text{do}(s_{<t}, \bar{s}_t, s'_{>t})$.

We begin with the definition of PNS as a counterfactual joint:

$$\text{PNS}(\mathbf{S}, \bar{s}_t) = P(\mathbf{A}_{\mathbf{S}} = \mathbf{y}, \mathbf{A}_{\mathbf{S}'} \neq \mathbf{y}).$$

Using the fact that $\mathbf{A}_{\mathbf{S}} = \mathbf{y}$ implies either $\mathbf{A}_{\mathbf{S}'} = \mathbf{y}$ or $\mathbf{A}_{\mathbf{S}'} \neq \mathbf{y}$, we can rewrite:

$$P(\mathbf{A}_{\mathbf{S}} = \mathbf{y}) = P(\mathbf{A}_{\mathbf{S}} = \mathbf{y}, \mathbf{A}_{\mathbf{S}'} = \mathbf{y}) + P(\mathbf{A}_{\mathbf{S}} = \mathbf{y}, \mathbf{A}_{\mathbf{S}'} \neq \mathbf{y}).$$

Thus:

$$\text{PNS}(\mathbf{S}, \bar{s}_t) = P(\mathbf{A}_{\mathbf{S}} = \mathbf{y}) - P(\mathbf{A}_{\mathbf{S}} = \mathbf{y}, \mathbf{A}_{\mathbf{S}'} = \mathbf{y}).$$

Under the Monotonicity assumption (Assumption 1), $\mathbf{A}_{\mathbf{S}'} = \mathbf{y} \Rightarrow \mathbf{A}_{\mathbf{S}} = \mathbf{y}$. This implies that the event $(\mathbf{A}_{\mathbf{S}} = \mathbf{y} \text{ and } \mathbf{A}_{\mathbf{S}'} = \mathbf{y})$ is equivalent to the event $(\mathbf{A}_{\mathbf{S}'} = \mathbf{y})$. Therefore:

$$P(\mathbf{A}_{\mathbf{S}} = \mathbf{y}, \mathbf{A}_{\mathbf{S}'} = \mathbf{y}) = P(\mathbf{A}_{\mathbf{S}'} = \mathbf{y}).$$

Hence:

$$\text{PNS}(\mathbf{S}, \bar{s}_t) = P(\mathbf{A}_{\mathbf{S}} = \mathbf{y}) - P(\mathbf{A}_{\mathbf{S}'} = \mathbf{y}).$$

Finally, under the Exogeneity assumption (Definition A.1), these counterfactual probabilities can be identified with interventional probabilities:

$$\text{PNS}(\mathbf{S}, \bar{s}_t) = P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S})) - P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S}')).$$

A.3 Identifiability of PNS without Monotonicity Assumption

Lemma 3 (Identifiability of PNS under $P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S})) = 1$ without Monotonicity)

Assume:

1. Under the perfect intervention of the correct CoT chain $\mathbf{S} = (\mathbf{s}_{<t}, \mathbf{s}_t, \mathbf{s}_{>t})$, the model always produces the correct answer:

$$P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S})) = 1.$$

2. Replacing step t with an incorrect step \bar{s}_t and allowing arbitrary rollout continuation $\mathbf{s}'_{>t}$ yields the intervened chain $\mathbf{S}' = (\mathbf{s}_{<t}, \bar{s}_t, \mathbf{s}'_{>t})$.
3. We do not assume Monotonicity (Assumption 1), i.e., we make no assumption that $P(\mathbf{A}_{\text{do}(\mathbf{S})} \neq \mathbf{y}, \mathbf{A}_{\text{do}(\mathbf{S}') = \mathbf{y}}) = 0$. (Note: original text had $\mathbf{A}_{\text{do}(\bar{s}_t)}$ which seems less precise here than $\mathbf{A}_{\text{do}(\mathbf{S}')}$ for the full chain).

Then the counterfactual joint defining PNS,

$$\text{PNS}(\mathbf{S}, \bar{s}_t) = P(\mathbf{A}_{\text{do}(\mathbf{S})} = \mathbf{y}, \mathbf{A}_{\text{do}(\mathbf{S}') \neq \mathbf{y}}),$$

is identifiable and simplifies to:

$$\text{PNS}(\mathbf{S}, \bar{s}_t) = 1 - P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S}')).$$

Start from the definition:

$$\text{PNS}(\mathbf{S}, \bar{s}_t) = P(\mathbf{A}_{\text{do}(\mathbf{S})} = \mathbf{y}, \mathbf{A}_{\text{do}(\mathbf{S}') \neq \mathbf{y}}).$$

Apply the law of total probability to the event $\mathbf{A}_{\text{do}(\mathbf{S})} = \mathbf{y}$:

$$P(\mathbf{A}_{\text{do}(\mathbf{S})} = \mathbf{y}) = P(\mathbf{A}_{\text{do}(\mathbf{S})} = \mathbf{y}, \mathbf{A}_{\text{do}(\mathbf{S}') \neq \mathbf{y}}) + P(\mathbf{A}_{\text{do}(\mathbf{S})} = \mathbf{y}, \mathbf{A}_{\text{do}(\mathbf{S}') = \mathbf{y}}).$$

Since $P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S})) = 1$ by assumption 1, the left side equals 1, so

$$1 = \text{PNS}(\mathbf{S}, \bar{s}_t) + P(\mathbf{A}_{\text{do}(\mathbf{S}') = \mathbf{y}}, \mathbf{A}_{\text{do}(\mathbf{S})} = \mathbf{y}).$$

Under Exogeneity (Definition A.1, implying no hidden confounding between the choice of intervention and its outcome), observing the original chain \mathbf{S} does not influence the outcome of intervening with \mathbf{S}' . Thus:

$$P(\mathbf{A}_{\text{do}(\mathbf{S}') = \mathbf{y}}, \mathbf{A}_{\text{do}(\mathbf{S})} = \mathbf{y}) = P(\mathbf{A}_{\text{do}(\mathbf{S}') = \mathbf{y}}) = P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S}')).$$

Substitute back to obtain

$$1 = \text{PNS}(\mathbf{S}, \bar{s}_t) + P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S}')),$$

and therefore

$$\text{PNS}(\mathbf{S}, \bar{s}_t) = 1 - P(\mathbf{A} = \mathbf{y} \mid \text{do}(\mathbf{S}')).$$

A.4 Equivalence of Perfect Intervention and Full Sufficiency

Theorem 1 (Equivalence of Perfect Intervention and Full Sufficiency) *Under the standard Exogeneity assumption (Definition A.1), and assuming there exists at least one alternative chain \bar{S} (which could be the original chain before applying the correct steps S , or any other relevant baseline) such that it has a nonzero probability of leading to failure, i.e., $P(A \neq y, \bar{S}, q) > 0$, the following two statements are equivalent:*

1. $P(A = y | do(S), q) = 1$ (*Perfect Intervention with S guarantees y*).
2. $PS(S, q) = 1$, where $PS(S, q) := P(A_{do(S)} = y | A \neq y, \bar{S}, q)$ (*Full Sufficiency*).

(\Rightarrow) Assume $P(A = y | do(S), q) = 1$. This means the intervention $do(S)$ guarantees $A = y$ in all worlds compatible with q . Consider the definition of $PS(S, q)$:

$$PS(S, q) = P(A_{do(S)} = y | A \neq y, \bar{S}, q).$$

Given the condition $A \neq y, \bar{S}, q$, we evaluate the probability of $A_{do(S)} = y$. Since $P(A = y | do(S), q) = 1$, it follows that $A_{do(S)} = y$ holds universally under $do(S)$, including in those specific circumstances where $A \neq y$ would have occurred with \bar{S} . Therefore, $P(A_{do(S)} = y | A \neq y, \bar{S}, q) = 1$, so $PS(S, q) = 1$.

(\Leftarrow) Assume $PS(S, q) = 1$. By definition, this means:

$$P(A_{do(S)} = y | A \neq y, \bar{S}, q) = 1.$$

This implies that for any world compatible with q where $A \neq y$ would occur with \bar{S} , applying $do(S)$ results in $A = y$. We want to show $P(A = y | do(S), q) = 1$, which is $P(A_{do(S)} = y | q) = 1$. Consider the outcome $A_{do(S)}$ given q . The intervention $do(S)$ sets the chain of thought to S and determines the outcome A . This outcome $A_{do(S)}$ is determined solely by S and q (due to exogeneity of S with respect to other factors once $do(S)$ is applied). If $PS(S, q) = 1$, it means $do(S)$ corrects all instances where \bar{S} would lead to failure. What about instances where \bar{S} might lead to success ($A = y$)? Since S is the “correct” chain of thought designed to produce y , the intervention $do(S)$ is assumed to robustly produce y . If it produces y when \bar{S} would have failed, and it (by its nature as a correct CoT) produces y when \bar{S} might have succeeded, then $A_{do(S)} = y$ holds across all situations defined by q and any alternative \bar{S} . Therefore, $P(A_{do(S)} = y | q) = 1$.

B Intervention Prompts

System Message for LLM Intervention

You are a helpful assistant. Continue solving the problem using mathematical expressions only, without repeating previous steps. Provide the final answer once, directly linked to the preceding reasoning, without additional summaries or explanations. Avoid using summarizing words such as ‘so’ or ‘thus,’ and refrain from repeating the final result when the calculation is already clear. Don’t say something like “Let’s continue with the previous reasoning” or other nonsense, just output the following reasoning directly.

Direct and External Intervention Prompt

Question:
{query}
Current reasoning steps:
{context_steps}

Prompt-Based Intervention Prompt

Ensure the next output node does not match the meaning of:
 {current_step}
 Avoid repeating the final result directly when the calculation is already clear.
 Question:
 {query}
 Current reasoning steps:
 {context_steps}

C In-Context Learning Prompts

ICL Baselines. The ICL baselines consist of the following variants:

1. **Standard:** This baseline uses verbose Chain-of-Thought (CoT) exemplars that include all intermediate reasoning steps, regardless of redundancy. It reflects the default strategy often employed in prompting LLMs for step-by-step reasoning.
2. **Fast-Solve:** This baseline encourages the model to produce concise reasoning chains that contain only the minimal steps necessary to reach the correct answer, avoiding verbose or redundant elaboration.
3. **Reduction** [19]: This method emphasizes rapid completion by prompting models to directly output shortcut solutions, often skipping step-by-step logical progression. It reflects a minimalistic strategy that favors brevity over transparency.
4. **Chain-of-Draft (CoD)** [67]: This variant uses prompts composed of minimally informative intermediate phrases—enough to scaffold the reasoning process but without detailed elaboration—simulating a rough-draft-style reasoning chain.
5. **Ours-ICL:** Our method, which leverages traces optimized using the *Probability of Necessity and Sufficiency* (PNS), presents only causally essential reasoning steps. These exemplars are pruned to retain only those steps that significantly contribute to correct outcomes, ensuring both efficiency and fidelity in reasoning.

System Message (Common for All Prompts)

You are a helpful assistant who is good at reasoning. Whenever doing multistep reasoning, please use two newline characters to split multiple steps (\n\n).

Ours-ICL Prompt (Sufficient and Necessary Reasoning)

User Message:

Instructions

When solving the following questions, your reasoning should:

- **Be Accurate:** Ensure your chain of thought leads to the correct answer without skipping any necessary logical steps.

- **Be Efficient:** Avoid unnecessary or redundant steps. Each step should be necessary to progress toward the solution.

- **Aim for Sufficient and Necessary Reasoning:** Only include steps that are both sufficient to reach the correct answer and necessary to avoid gaps or confusion. If a step can be removed without affecting correctness, remove it.

- **Notice the Pattern:** In the following examples, compare the original, verbose solution with the optimized solution. Learn to identify and eliminate redundant reasoning steps while preserving logical soundness.

Example 1: {example1}

Example 2: {example2}

Example 3: {example3}

Now Solve This:

Question: {question}

Your Simplified and Optimized Answer:

Fast-Solve Prompt

User Message:

You are a math assistant that solves problems step by step. Please reason in a clear and structured manner, but keep your explanation as concise as possible. Avoid unnecessary repetition or redundant steps. The goal is to solve the problem accurately with the fewest necessary steps.

Now Solve This:

Question: {question}

Your Simplified and Optimized Answer:

CoD Prompt

User Message:

Think step by step, but only keep a minimum draft for each thinking step, with 5 words at most. Return the answer at the end of the response after a separator

Now Solve This:

Question: {question}

Your Simplified and Optimized Answer:

Reduction Prompt

User Message:

Let's quickly conclude the answer with shortcut reasoning.

Now Solve This:

Question: {question}

Your Simplified and Optimized Answer:

D Supervised Fine-Tuning (SFT) Hyperparameter Settings

Table 4 details the general hyperparameter configuration used for supervised fine-tuning (SFT) in our experiments. All SFT training was conducted on 8 NVIDIA RTX 3090 GPUs using the ZeRO-3 optimizer for efficient memory distribution. To accelerate training and reduce memory usage, we employed bf16 mixed-precision computation.

The training used the `flash_attention_2` implementation for efficient attention computation, combined with a cosine learning rate scheduler that decays to a minimum learning rate. Each GPU was assigned a batch size of 1 due to the large context length of 16,384 tokens. The model was trained for 3 epochs, and `max_steps` was left as -1 to allow epoch-based termination. These settings balance computational feasibility and performance under long-context, reasoning-intensive tasks.

The same configuration was applied across all target models, including DeepSeek-R1-Qwen-1.5B, DeepScaleR-1.5B-Preview, and Phi-4-mini-reasoning, unless otherwise specified.

Table 4: General SFT Hyperparameters. Hardware: $8 \times$ NVIDIA RTX 3090 GPUs, ZeRO-3 optimizer, bf16 mixed precision.

Parameter	Value
attn_implementation	flash_attention_2
bf16	true
learning_rate	5.0e-05
lr_scheduler_type	cosine_with_min_lr
per_device_train_batch_size	1
max_steps	-1
max_length	16384
num_train_epochs	3

E In-Context Learning: Case study

To further illustrate the effectiveness of our optimized CoT examples in in-context learning (ICL), we provide a case study using the Qwen-2.5-72B-Instruct model on a representative problem from the MATH-500 dataset.

Figure 4 compares two responses: one directly generated by the Qwen-2.5-72B-Instruct model without any in-context examples (blue background), and the other generated under ICL using our optimized CoT example (pink background).

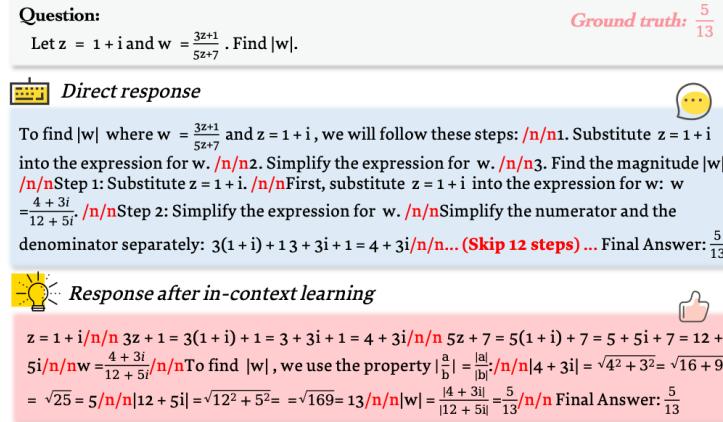


Figure 4: Case Study: Comparison of direct response from Qwen-2.5-72B-Instruct (blue background) and response under ICL with optimized CoT examples (pink background) on a MATH-500 problem. The optimized CoT enables more sufficient and necessary reasoning.

The direct response exhibits a lengthy reasoning process with several redundant or unnecessary steps and expressions. In contrast, the ICL-guided response is more concise and logically structured, reflecting a clearer and more efficient problem-solving strategy.

This comparison demonstrates how our optimized CoT exemplars help guide the model toward more focused and causally sufficient and necessary reasoning.

F Full Version of Related Work

F.1 Reasoning Sufficiency Enhancement via CoT Optimization.

Recent efforts have focused on improving the reasoning capabilities of large language models (LLMs) through the development of Chain-of-Thought (CoT) reasoning and its variants. CoT [61] introduced intermediate reasoning steps to enable LLMs to perform structured, multi-step reasoning. This foundational idea has since evolved into more sophisticated frameworks such as Tree-of-Thought (ToT)[76] and Graph-of-Thought (GoT)[6], which organize reasoning structures into tree and graph

forms, respectively. Yue et al. [78] introduce DOTS, a method enabling dynamic reasoning trajectory planning via optimal reasoning strategy search, resulting in more adaptive and efficient problem-solving. Jin et al. [28] introduce a CoT framework using graph structures for iterative reasoning in LLMs and builds the GRBench dataset for graph-based reasoning evaluation. Yang et al. [69] explores how CoT prompts affect LLM mechanisms, enhancing knowledge retrieval by activating more neurons.

Several recent innovations further expand the reasoning capacity of LLMs. LLaVA-CoT [65] combines CoT with a multimodal visual-language model to enhance reasoning in vision-language tasks. Meta-CoT [64] formulates reasoning as a latent variable process, improving flexibility and generalization. Markov Chain of Thought [73] introduces Markovian transitions across reasoning steps by clearing the context KV cache to extend reasoning depth. Additionally, Puerto et al. [47] propose a self-correction mechanism using multiple reasoning paths, significantly improving performance on knowledge-intensive benchmarks. H  y  k et al. [26] propose fine-tuning strategies using counterfactual feedback to enhance LLMs' causal reasoning capabilities. Ma et al. [38] demonstrate that explicit thinking processes are not always necessary and propose a simplified "NoThinking" method achieving competitive reasoning performance with reduced computational costs. Diao et al. [18] present Active-Prompt, a method that enhances LLM reasoning by selectively annotating task-specific prompts for automatic adaptation. ECHO [31] unifies diverse reasoning paths to improve the consistency and accuracy of LLM reasoning.

Additional recent studies have explored novel dimensions in CoT reasoning. Ding et al. [19] propose "Break the Chain" strategies, integrating heuristic shortcuts to streamline CoT reasoning, significantly enhancing efficiency. Stolfo et al. [51] identify entropy and token frequency neurons, elucidating internal mechanisms by which LLMs manage uncertainty and confidence. Ali et al. [2] mitigate copy biases in in-context learning through targeted neuron pruning, improving generalization. De Sabbata et al. [15] use rational metareasoning to selectively invoke intermediate steps, reducing inference cost while preserving accuracy. Turpin et al. [58] question CoT's faithfulness by systematically evaluating model-generated rationales. Jin et al. [30] demonstrate that artificially lengthening reasoning can superficially boost performance, highlighting potential pitfalls in CoT evaluation. Simhi et al. [49] reveal that LLMs can exhibit high-certainty hallucinations, producing incorrect answers with strong confidence, thereby challenging the reliability of CoT-based outputs.

Despite their benefits, these structured reasoning methods often introduce excessive token length, which becomes problematic in cost-sensitive or latency-constrained scenarios [62]. Moreover, models frequently fail to assess task complexity, leading to over-reasoning on simple problems—an issue known as overthinking [12, 7, 36]. Our proposed approach addresses these limitations by simultaneously ensuring reasoning efficiency and maintaining high accuracy.

F.2 Reasoning CoT Redundancy

Recent studies on CoT redundancy have sought to mitigate redundancy in CoT reasoning. Token-budget-aware methods, such as Han et al. [22], dynamically allocate reasoning budgets based on task complexity. C3oT [32] leverages GPT-4 as a compressor to retain only essential reasoning content. CCoT [9] and COCONUT [23] adopt continuous representations to encode reasoning traces more compactly, while CoT-Valve [39] introduces variable-length CoTs. Training-free approaches, including Kimi K1.5 [54] and O1-Pruner [36], use prompt ensembles or reinforcement learning to discard unnecessary reasoning steps.

Recent efforts also investigate finer-grained control over CoT reasoning. Wu et al. [63] introduces the Thinking Intervention paradigm, allowing targeted interventions on reasoning tokens. Yang et al. [74] highlight the impact of excessive CoT length on model performance, proposing adaptive token scaling strategies. Liu et al. [35] demonstrate external thoughts from smaller models can effectively streamline reasoning in larger models, reducing redundancy.

However, most of these methods focus primarily on sequence length reduction or representation compression, without explicitly addressing causal logical redundancy. In contrast, our approach identifies and preserves only those reasoning steps that are both causally sufficient and necessary, grounded in a formal intervention-based framework. This results in CoT traces that are not only more compact but also more causally meaningful.

F.3 PNS Theory in CoT Reasoning.

Our work is grounded in the causal inference framework proposed by Pearl [46], which defines the Probability of Sufficiency (PS) and the Probability of Necessity (PN) to quantify whether a cause is sufficient or necessary for a given effect. For instance, in image classification, PS measures how likely adding a feature (e.g., “pointy ears”) leads to a positive label (e.g., “cat”), while PN measures how likely removing that feature would change the outcome.

While Hüyük et al. [26] introduced PN and PS as evaluation metrics to enhance model-level causal reasoning via counterfactual fine-tuning, our approach advances this direction by applying these metrics to the internal reasoning traces of Chain-of-Thought (CoT) prompting. Specifically, we formalize and operationalize the Probability of Necessary and Sufficient Causes (PNS) for individual reasoning steps, enabling a structured intervention-based analysis that identifies steps that are both causally essential and logically minimal.

Unlike prior work that emphasizes model-wide causal consistency, our framework targets step-level causal minimality within CoT, yielding concise, interpretable, and causally grounded reasoning sequences. This design allows us to optimize CoTs not merely for brevity or accuracy, but for causal soundness. Crucially, our method is model-agnostic and applies to any LLM capable of generating CoT outputs, representing a significant advancement over heuristic or model-specific compression strategies.

Our approach is essentially a causal analysis method for multi-step systems. In fact, it is not only applicable to Chain-of-Thought (CoT) reasoning, but can also inspire causal thinking in computational experiments [68, 77] and other complex systems.

G Additional PNS Comparison Results

This appendix provides comprehensive visualizations of average PNS values before and after applying our optimization algorithm across multiple models and datasets. Each figure presents results over 30 sampled problems, highlighting the increased causal necessity of the retained reasoning steps after optimization. See Figures 5 to 8 for detailed comparisons.

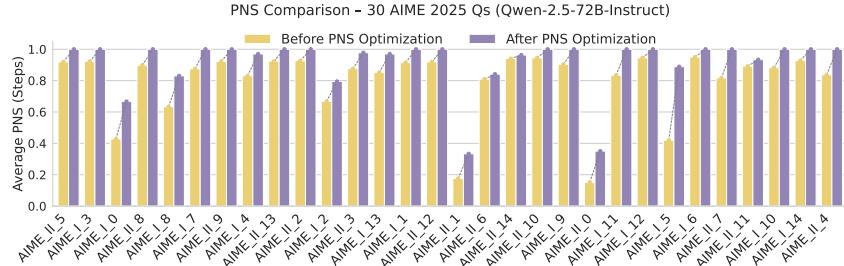


Figure 5: Average PNS comparison for Qwen-2.5-72B-Instruct on the AIME dataset.

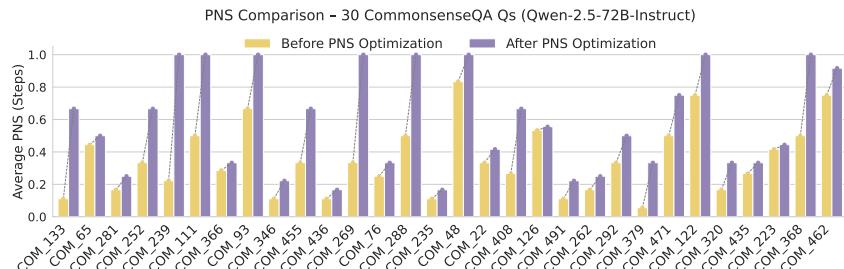


Figure 6: Average PNS comparison for Qwen-2.5-72B-Instruct on the CommonsenseQA dataset.

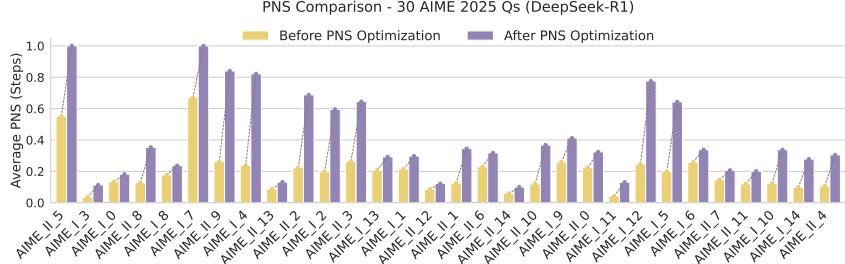


Figure 7: Average PNS comparison for DeepSeek-R1 on the AIME dataset.

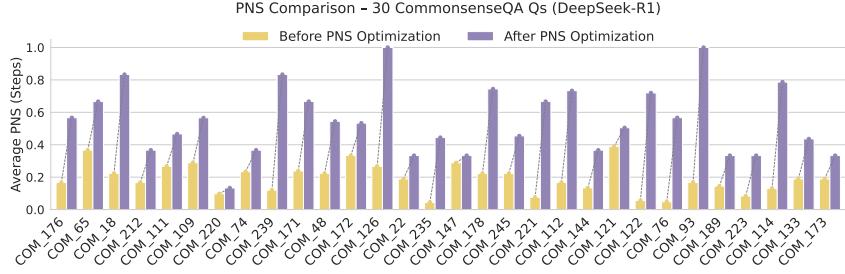


Figure 8: Average PNS comparison for DeepSeek-R1 on the CommonsenseQA dataset.

H Computational Complexity Analysis

Let us denote the following quantities:

- n : the number of reasoning steps in the chain-of-thought (CoT);
- k : the number of rollouts per step;
- l_{step} : the average number of tokens generated per step;
- l_{out} : the number of tokens in each verification output (assumed constant);
- t_{token} : the time required to generate a single token (assumed constant).

We analyze the computational cost of our PNS-based intervention method by separately evaluating the *rollout* and *evaluation* stages.

H.1 Rollout Time Complexity

At each reasoning step $i \in \{1, \dots, n\}$, the model performs k rollouts, where each rollout generates the remaining $(n - i)$ steps. Each step contains on average l_{step} tokens. Thus, the total rollout time is given by:

$$T_{\text{rollout}} = \sum_{i=1}^n k \cdot (n - i) \cdot l_{\text{step}} \cdot t_{\text{token}} = O(k \cdot l_{\text{step}} \cdot t_{\text{token}} \cdot n^2) \quad (6)$$

H.2 Evaluation Time Complexity

Each rollout must also be evaluated to determine its validity. This involves generating both the full continuation of the chain and a final answer output of length l_{out} . The total evaluation time is therefore:

$$T_{\text{eval}} = \sum_{i=1}^n k \cdot [(n - i) \cdot l_{\text{step}} + l_{\text{out}}] \cdot t_{\text{token}} = O(k \cdot t_{\text{token}} \cdot (l_{\text{step}} \cdot n^2 + l_{\text{out}} \cdot n)) \quad (7)$$

Since typically $l_{\text{out}} \ll l_{\text{step}} \cdot n$, the second term is asymptotically negligible, yielding:

$$T_{\text{eval}} = O(k \cdot l_{\text{step}} \cdot t_{\text{token}} \cdot n^2) \quad (8)$$

H.3 Total Complexity

Combining the rollout and evaluation phases, the total computational cost becomes:

$$T_{\text{total}} = T_{\text{rollout}} + T_{\text{eval}} = O(k \cdot l_{\text{step}} \cdot t_{\text{token}} \cdot n^2) \quad (9)$$

Assuming that both l_{step} and t_{token} are constants (or near-constants) in practice, the overall time complexity simplifies to:

$$T_{\text{PNS}} = O(k \cdot n^2) \quad (10)$$

This quadratic dependence on n highlights the computational cost of deeper reasoning chains, while the linear dependence on k reflects the tradeoff between rollout breadth and computation time.

I Qualitative Analysis of Whether the Final Reasoning Is Sufficient and Necessary

We conducted a human evaluation of 50 chain-of-thought (CoT) samples generated by a model fine-tuned on data curated with our PNS-based algorithm. Five mathematics experts independently judged each sample under three labels:

- **S&N**: reasoning is both sufficient and necessary to support the final answer;
- **SbU**: reasoning is sufficient but contains unnecessary (redundant) steps;
- **NbI**: reasoning is insufficient (missing critical steps).

Table 5: Human evaluation of reasoning quality. “Fully Sufficient” = S&N + SbU. “Redundant” flags any chain containing redundant steps and may co-occur with NbI.

Dataset	# Samples	Fully Sufficient	Redundant	S&N	SbU	NbI
GSM8K	20	19	3	17	2	1
Commonsense QA	15	15	1	14	1	0
MATH500	15	13	4	11	2	2
Total	50	47 (94.0%)	8 (16.0%)	42 (84.0%)	5 (10.0%)	3 (6.0%)

Most outputs are logically sound under the sufficiency/necessity criterion: 84% of CoTs are both sufficient and necessary, and only 6% (3/50) are insufficient. MATH500 exhibits a higher incidence of redundancy and incompleteness, consistent with the greater difficulty of mathematical reasoning in this benchmark.

J Validator Accuracy and Robustness

To assess the reliability of the PNS validator V , we evaluate its performance across different LLMs and rollout sizes k . Specifically, we report the *mean absolute error* (MAE) between the estimated PNS values and ground-truth labels, where ground-truth PNS is defined as 1 for reasoning chains judged by experts to be both necessary and sufficient.

Metric. The evaluation computes the mean absolute difference $|\widehat{\text{PNS}} - \text{PNS}|$ across annotated samples. Lower values indicate better alignment with ground truth.

We compare validator accuracy across rollout sizes k and model strengths. Increasing k and using stronger LLMs both reduce MAE, yielding more stable and accurate PNS estimates. GPT-4o achieves the lowest error across all settings, underscoring that validator quality and model strength are key to reliable counterfactual analysis.

Table 6: Validator MAE Across LLMs and Rollout Sizes. Lower is better.

Validator	$k=1$	$k=3$	$k=5$	$k=10$
Qwen-72B	0.315	0.187	0.142	0.116
Qwen-7B	0.411	0.395	0.315	0.293
GPT-4o	0.137	0.114	0.090	0.050

K Supplementary Experimental Results

To further assess the robustness and generality of our proposed method, we compare it with three representative reasoning baselines: **SPIRIT**, **ReAct**, and **Tree-of-Thoughts (ToT)**. These methods represent distinct paradigms for enhancing large language model reasoning:

- **SPIRIT** [13] uses perplexity to identify key reasoning steps and prune redundant tokens.
- **ReAct** [75] integrates reasoning and acting by alternating between *thought* and *action* steps, improving interpretability and interactive decision-making.
- **Tree-of-Thoughts (ToT)** [76] introduces a tree-structured reasoning process, maintaining multiple reasoning trajectories to perform deliberate exploration and self-evaluation. We implement this process using a simple prompt-based approach.

Table 7 summarizes the results across three benchmarks—*CommonsenseQA*, *GSM-8K*, and *MATH-500*. Each metric is reported as **Tokens \downarrow / Steps \downarrow / Acc. \uparrow** , where fewer tokens and steps indicate greater reasoning efficiency.

Table 7: Comparison of SPIRIT, ReAct, ToT, and Ours-ICL across models and datasets. Each metric is reported as **Tokens \downarrow / Steps \downarrow / Acc. \uparrow** .

Model	Method	CommonsenseQA			GSM-8k			MATH-500		
		Tokens \downarrow	Steps \downarrow	Acc. \uparrow	Tokens \downarrow	Steps \downarrow	Acc. \uparrow	Tokens \downarrow	Steps \downarrow	Acc. \uparrow
DeepSeek-V3	SPIRIT	142.3	4.4	83.9	73.1	2.0	95.5	250.5	7.7	89.6
	ReAct	181.2	5.4	83.8	179.5	6.9	92.3	412.3	19.0	91.2
	ToT	271.8	8.2	80.1	198.9	7.7	91.7	349.1	13.9	72.8
	Ours-ICL	44.7	2.7	83.6	52.2	4.3	99.9	136.7	6.4	96.2
Qwen-2.5-72B-Instruct	SPIRIT	214.2	7.6	76.3	98.3	4.0	94.2	239.3	10.3	81.4
	ReAct	175.0	4.1	83.7	171.7	4.1	92.5	241.6	7.9	77.6
	ToT	269.2	7.9	80.0	273.5	10.4	75.1	345.1	14.1	70.6
	Ours-ICL	81.6	3.4	83.0	65.3	5.3	99.5	196.9	8.9	81.2
Qwen-2.5-7B-Instruct	SPIRIT	188.6	6.5	59.3	113.7	5.1	86.7	238.8	9.4	72.9
	ReAct	192.9	5.2	79.4	181.6	7.2	84.5	263.0	10.5	70.8
	ToT	335.4	10.1	78.0	212.3	8.9	70.5	302.6	12.3	50.0
	Ours-ICL	99.1	3.8	77.6	83.4	4.8	94.1	174.7	7.7	72.6
LLaMA-3.1-8B-Instruct	SPIRIT	170.1	7.4	63.2	179.7	7.8	80.9	1218.1	67.3	42.8
	ReAct	382.1	17.9	79.0	521.5	20.6	82.5	1176.9	53.6	46.0
	ToT	559.9	19.4	75.3	711.7	27.1	69.0	996.7	45.8	30.2
	Ours-ICL	120.3	7.1	72.1	128.3	8.1	93.1	365.9	24.2	54.8

Across all models and datasets, our proposed **Ours-ICL** consistently achieves a superior balance between reasoning accuracy and efficiency. While ReAct and ToT obtain competitive accuracy on commonsense reasoning tasks, they dramatically increase reasoning cost in both tokens and steps. SPIRIT improves efficiency via perplexity-based step selection but can underperform on complex mathematical reasoning. In contrast, **Ours-ICL** yields the best overall trade-off, reducing reasoning cost by up to 60–80% while maintaining or improving accuracy, especially on *GSM-8K* and *MATH-500*.