Correlation or Causation: Analyzing the Causal Structures of LLM and LRM Reasoning Process

Zhizhang Fu*, Guangsheng Bao*, Hongbo Zhang, Chenkai Hu, Yue Zhang†, Member, IEEE

Abstract-LLMs suffer from critical reasoning issues such as unfaithfulness, bias, and inconsistency, since they lack robust causal underpinnings and may rely on superficial correlations rather than genuine understanding. Successive LRMs have emerged as a promising alternative, leveraging advanced training techniques such as reinforcement learning (RL) and distillation to improve task accuracy. However, the impact of these training methods on causality remains largely unexplored. In this study, we conduct a systematic causal analysis on LLMs and LRMs, examining structural causal models (SCMs) of four key variables: problem instruction (Z), thinking process (T), reasoning steps (X), and answer (Y). Our findings reveal that RLVR-trained LRMs exhibit enhanced causal reasoning capabilities, aligning more closely with ideal causal structures, while LLMs and distilled LRMs fail to address causalityrelated deficiencies. Our further investigation indicates that RLVR reduces spurious correlations and strengthens genuine causal patterns, thereby mitigating unfaithfulness and bias. In addition, our inspection on the dynamics of the RLVR training process observes a high correlation between reduced spurious features and improved causal structures, where the causal relationships consistently improve in the training process. This study contributes to the understanding of causality in reasoning models, highlights the critical role of RLVR in enhancing causal reasoning, and provides insights for designing future AI systems with stronger causal foundations. We release our code and data at https://github.com/Harryking1999/CoT_Causal_Analysis.

Index Terms—LLM reasoning, causality, reinforcement learning, distillation.

I. INTRODUCTION

ARGE Language Models (LLMs; e.g., GPT4 [1], DeepSeek-V3 [2]) and successive Large Reasoning Models (LRMs; e.g., o1 [3], DeepSeek-R1 [4]) have demonstrated impressive reasoning capabilities, producing explainable reasoning processes (chain of thoughts) [5]–[7]. LRMs differ from LLMs in further giving a thinking process before explicit reasoning. Despite showing effectiveness in math and other tasks, CoT reasoning processes often suffer from critical issues such as unfaithfulness, bias, and inconsistency [8]–[12], which has caused concerns about its reliability [13]–[15]. Empirical evidence shows that correct CoTs may lead to incorrect answers, and incorrect CoTs to correct answers [12].

Intuitively, robust reasoning is rooted in a causal chain of thoughts, not just superficial correlations between thoughts. An ideal reasoning process often involves identifying causal relationships to make sense of existing facts and to predict future occurrences [16], which leads to stronger robustness and generalization [17]. Therefore, analyzing LLM and LRM

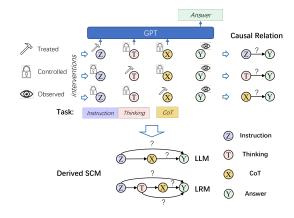


Fig. 1. Causal Analysis to derive underlying causal structure of LLMs and LRMs. We conduct treated experiments that interfere (treat) one variable while freeze (control) others, so that we can observe their effects.

reasoning from a causal perspective is essential for deeper insights into their behavior and for advancing the reasoning models, answering essential questions including *when* and *how* a causal reasoning process happens, which existing work does not consider [11], [18], [19].

We conduct a causal analysis for CoT, dividing the reasoning process into four parts, each represented by a random variable. Specifically, as Figure 1 illustrates, they are problem instruction (Z), thinking process (implicit CoT, T, for LRM only), reasoning steps (explicit CoT, X), and final conclusion (answer, Y). We infer the causal relationships between these variables and the answer Y, where a structural causal model (SCM) can be derived for the reasoning process. Typically, we identify four major types and seven minor types of causal structures, as Figure 2 shows. An ideal reasoning process is represented as a causal chain (type-I), where the instruction determines the reasoning, and the reasoning decides the answer. In this case, the model is faithfully reasoning, which produces consistent responses. In contrast, when a reasoning process has a common cause (type-II) structure, it is actually explaining a latent belief of the answer, which may produce unfaithful and inconsistent responses.

We evaluate various LLMs and LRMs on a range of reasoning tasks (Section IV-C). Experimental results show that CoT of LLMs generally does not possess an ideal causal structure. In contrast, LRMs generally achieve higher task accuracy compared to LLMs, but their causal structures vary significantly depending on the training method. Distilled LRMs [4], which rely on supervised fine-tuning (SFT) [20], [21], do not exhibit improved causal structures and often share similar

^{*}Contributed equally to this work as co-first authors

[†]Corresponding author: yue.zhang@wias.org.cn

2

causal deficiencies as LLMs. LRMs trained with reinforcement learning (i.e., reinforcement learning with verifiable rewards, RLVR) [22]–[26] demonstrate enhanced causality, aligning more closely with the ideal causal chain structure. These findings suggest that different LLM training paradigms may influence the causal reasoning capabilities differently.

We analyze various techniques that could potentially influence the causal structures, including in-context learning (ICL), instruction-tuning, reinforcement learning with human feedback (RLHF), long CoT distillation, and RLVR (Section VI-B). We find that RLVR substantially enhances causal structures, whereas other techniques have a minimal impact. Specifically, ICL could strengthen the causal relationships with limited effects, while instruction-tuning, RLHF, and distillation weaken the causal structures. Among all the techniques, RLVR is the strongest for achieving ideal reasoning structures.

To further understand the impact of RLVR on causality, we investigate the dynamics of causal structure evolution during RLVR training (Section VI). Our experiments show that RLVR consistently improves the causal relationships over the course of training, by reducing spurious correlations and strengthening models' genuine causal patterns, leading to more robust reasoning models. This dynamic improvement highlights the potential of RLVR as a powerful tool for enhancing causality in reasoning models, addressing key issues such as unfaithfulness, bias, and inconsistency.

This manuscript significantly extends our conference paper [12] which focuses on the causality of LLM CoT. First, we extend causal analysis from LLMs to LRMs, revealing that LRMs exhibit enhanced causal reasoning capabilities. Second, we systematically investigate various learning paradigms, finding that RLVR substantially enhances causal structures. Finally, we discuss why RLVR improves causality by conducting extensive empirical experiments. These investigations and findings offer valuable insights for the development of more reliable and causally robust reasoning models. Our findings pave the way for future research on causality-driven LLM systems and highlight the importance of RL in advancing reasoning capabilities.

II. RELATED WORK

Various reasoning techniques have been proposed to enhance the reasoning ability of LLMs [27], [28]. Chain-of-thought (CoT) prompting [5], as an early study elicits reasoning in LLMs, inspires numerous further investigations. Specifically, self-consistency [29] votes the major decision from multiple reasoning paths, Tree-of-thought [30] searches the most confident reasoning path in a tree, and Graph-of-thought [31] represents the thoughts as graph nodes and combines thoughts non-sequentially. Advanced CoT methods, like Faithful CoT [32] and Constraint CoT [33], are further proposed to improve reasoning capabilities. In this paper, we focus on the very basic chain of thought to understand the underlying mechanism of how LLMs do reasoning, leaving the analysis of advanced methods for the future.

Recently, researchers have introduced Large Reasoning Models (LRMs) specifically designed for reasoning, achieving

significant progress on reasoning problems. Leading models, including Claude 4 [34], DeepSeek-R1 [4], Gemini 2.5 Pro [35], Qwen 3 [36], Grok 4 [37], OpenAI o3 [38] and GPT-5 [39], now set the pace on reasoning benchmarks like MMLU-PRO [40], GPQA [41] and AIME [42]. Collectively, these results indicate rapid progress toward more reliable general reasoning competence in diverse tasks. However, the reason for LRMs achieving stronger results in reasoning has not been understood from a causality perspective. Our work fills this gap.

Among the LRM methods, OpenAI initially leverages large-scale reinforcement learning [3], marking the beginning of the LRM era. Following this breakthrough, the community has conducted extensive researches to reproduce o1-like LRMs, typically employing various RL algorithms for training and Monte Carlo Tree Search for test-time scaling [43]–[45]. Since the release of DeepSeek-R1 [4], Reinforcement Learning with Verifiable Rewards (RLVR) has garnered considerable attention, particularly through the Group Relative Policy Optimization (GRPO) algorithm [22]. The efficacy of GRPO in enhancing reasoning capabilities has spawned several variants, including Group Sequence Policy Optimization (GSPO) implemented in Qwen models [23], DAPO [24], Dr.GRPO [25], and GPG [26]. RLVR has become a prominent post-training methodology for training LRMs.

Meanwhile, supervised fine-tuning (SFT) represents another prevalent training paradigm for LRMs. Muennighoff et al. [46] demonstrate that SFT with a small amount of high-quality data can achieve superior performance compared to o1-preview [3] on mathematical benchmarks. Concurrently, distillation has emerged as a mainstream SFT approach in recent LRM training. This method leverages the outputs of larger-scale models to empower smaller models with reasoning capabilities [4], enabling dramatic improvements in reasoning performance that approach the levels of their larger counterparts.

Despite the remarkable achievements of LRMs, some studies have questioned whether they engage in genuine reasoning. For example, Shojaee et al. [47] demonstrate that LRMs exhibit advantages in medium-complexity tasks, underperform compared to standard LLMs in low-complexity tasks, and both model types experience complete collapse when faced with high-complexity tasks. Additionally, there is considerable debate regarding the relative merits of the two primary training paradigms for LRMs: RL and distillation. On the one hand, RL demonstrates unparalleled advantages over other paradigms in training large-scale LRMs; on the other hand, for smaller and less capable models, only distillation can introduce new knowledge and rapidly improve performance, whereas RL can merely enable models to better utilize their existing knowledge, being constrained by the original model's knowledge limitations [48]. In this work, we analyze the relative merits of distillation and RL from a causal perspective.

Causality represents the fundamental building blocks of both physical reality and human understanding, transcending mere statistical associations to capture the underlying mechanisms that govern phenomena. Moreover, causal relationships exhibit greater robustness and stability than probabilistic relationships, remaining "invariant to changes in all mechanisms" and pro-

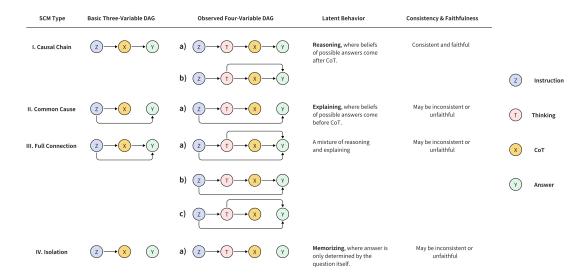


Fig. 2. Structural causal model (SCM) types and their latent behavior, consistency, and faithfulness. We use the three-variable DAGs for LLMs while the four-variable DAGs for LRMs.

viding the foundation for genuine understanding rather than mere pattern recognition [16]. This highlights the importance of performing causal assessment on current LLMs. However, existing studies on the causal reasoning capabilities of LLM mainly focus on variables described in natural language, like benchmark ATOMIC [49], CLadder [50], and Corr2Cause [51]. ATOMIC focuses on the if-then relations of variables like "if X pays Y, Y will probably return". CLadder requires the identification of variables and their causal relations from the language context prior to inference. Corr2Cause determines the causal structure between variables according to a group of correlational statements. On multiple causal benchmarks, [52] finds that LLMs achieve good accuracy and hypothesizes that LLMs can use their collected knowledge to generate causal graphs from natural language, while [53] conjectures that a successful causal inference relies on a pre-learned meta-SCM that stores the related causal facts in natural language. Unlike these studies, where variables represent targets in the question domain, we investigate the causality between four known variables, instruction, CoT, thinking and answer, which do not represent any question-specific target, but only abstractive components of the chain-of-thought reasoning.

III. CAUSAL ANALYSIS

As shown in Figure 1, we take an intervention-based approach to conduct causal analysis for CoT, where a causal graph is constructed (Section III-A), and causal relations are calculated (Section III-B), so that the CoT patterns can be categorized into causal structure prototypes (Section III-C), and different prototypes exhibit distinct CoT behaviors (Section III-D).

A. Random Variables

We model the reasoning process of LLMs in three random variables and LRMs in four random variables, involving instruction (Z), CoT (X) and answer (Y) for LLMs and an additional thinking (T) for LRMs, described as follows.

Instruction (*Z*) generally consists of a task description, a question formulation, several examples, to guide the models toward an appropriate response. Because few-shot prompting is not recommended in some reasoning models [4], our instructions follow the simplest style, directly describing the question with a zero-shot setting.

Thinking (T; for LRM only) is a new component introduced in LRMs. It consists of a long chain-of-thought (CoT), characterized by extensive exploration, feasible reflection, and deep reasoning [54]. This stage provides an internal process by which the model thoroughly analyzes the problem before generating the explicit CoT or answer. In this work, we investigate whether the inclusion of the thinking stage has any impact on the overall causality of reasoning in the models.

CoT (X) is the step-by-step reasoning process of LLMs, which is often questioned for lacking faithfulness. We therefore set CoT as an independent variable and verify its causal relationship with Answer.

Answer (*Y*) represents the final step of the reasoning process, which answers the question. Unlike the two variables above, which work as either independent variables or constants, answer consistently serves as a dependent variable in experiments. All experiments focus on studying the relationships between other variables and the answer, which enables us to evaluate the reasoning causality structures.

B. Identification of Causal Relations

In autoregressive language models, a token can depend on any tokens on the left, therefore presenting arbitrary subgraphs of fully connected structural causal models (SCMs). To figure out the specific SCM a model follows, we employ treatment experiments [55], [56], with each time intervening only one independent variable and observing answer (Y) as the dependent variable. Below are several key definitions:

Definition 3.1 (Cause-Effect Interventions): Suppose that the SCM \mathcal{G} entails a distribution $P_{X,Y}$ with $N_X, N_Y \stackrel{iid}{\sim}$

 $\mathcal{N}(0,1).$ Then we intervene on X to change the distribution of Y

$$P_Y^{do(X)} = P(Y|do(X)). \tag{1}$$

By intervening on an independent variable and observing the dependent variable before and after the intervention, we can determine whether they are causally related.

Definition 3.2 (Average Treatment Effect; ATE): An ATE [55] represents the effect of an intervention, which compares the distributions of the target variable Y with and without a treatment.

$$ATE = E(Y|do(X)) - E(Y).$$
 (2)

We use the significance of ATE to determine the causal relationship between two variables. Specifically, we employ McNemar's test [57] for assessing the significance. Since our interventions are all negative, we take the absolute value of negative ATEs and set positive ATEs to zero, in order to avoid confounding the causal relationship.

Definition 3.3 (Relative Average Treatment Effect; R-ATE): We introduce R-ATE to measure the proportion of the actual treatment effect relative to the maximum possible treatment effect. It indicates what percentage of the total achievable effect is actually realized.

$$R-ATE = \frac{ATE}{Possible Max ATE}$$
 (3)

The value of ATE is not only influenced by underlying causal mechanisms we aim to measure, but also by other irrelevant factors. For instance, if model A changes from 0.9 to 0, and model B changes from 0.7 to 0 under the same intervention, both are completely intervened, but their ATEs differ, making ATE incomparable across different models. R-ATE addresses this issue by enabling fair comparisons across models. However, R-ATE is still not comparable across different intervention methods.

Particularly in our experiments, we test different hypotheses depending on the SCM structure. For three-variable SCMs, we test two hypotheses: whether instruction causes the answer, and whether CoT causes the answer. For four-variable SCMs, we test **three hypotheses**: whether instruction, thinking and CoT, causes the answer.

Hypothesis 3.1 (Instruction causes Answer): Given a constant CoT X and Thinking T

$$\left\{ \begin{array}{ll} H_0: \text{ATE} = 0, & \text{Instruction does not cause Answer,} \\ H_1: \text{ATE} \neq 0, & \text{Instruction causes Answer,} \end{array} \right.$$

where ATE = E(Y|T, X, do(Z)) - E(Y|T, X).

Hypothesis 3.2 (CoT causes Answer): Given a constant Instruction Z and Thinking T

where ATE = E(Y|Z, T, do(X)) - E(Y|Z, T).

Hypothesis 3.3 (Thinking causes Answer): Given a constant Instruction Z and CoT X

$$\left\{ \begin{array}{l} H_0: {\rm ATE}=0, \quad {\rm Thinking\ does\ not\ cause\ Answer}, \\ H_1: {\rm ATE}\neq 0, \quad {\rm Thinking\ causes\ Answer}, \end{array} \right. \eqno(6)$$

where ATE = E(Y|Z, X, do(T)) - E(Y|Z, X).

To validate these hypotheses, we employ distinct intervention strategies for each. For Hypothesis 3.2 and Hypothesis 3.3, we intervene on the CoT by replacing it with randomized reasoning chains. For Hypothesis 3.1, we intervene on the instruction with biases like "I think the answer is w", where w represents an incorrect answer. The statistical significance of these hypotheses reveals the underlying SCM of each LLM in each task, and the R-ATE provides more finegrained information as a continuous value rather than a binary classification like significance.

C. SCM Types

As shown in Figure 2, we define the four basic SCM types using three-variable directed acyclic graphs (DAGs). To maintain consistency and comparability, we align the four-variable SCMs with the three-variable SCMs by treating the combined effects of T and T on T as equivalent to the effect of T on T in the three-variable case (where significance of either T or T on T corresponds to significance of T on T0, while the effect of T2 on T1 remains identical across both frameworks. Therefore, the four-variable SCMs and three-variable SCMs capture the same underlying causal relationships, with the four-variable SCMs providing a finer-grained representation.

Type I. Causal Chain is the ideal SCM in our experimental setting, where the answer is caused by CoT and not by Instruction. This represents the desired behavior in which the model derives answers based on variables that should have a causal effect.

Type II. Common Cause signifies an SCM that Instruction causes the answer while CoT does not. In this case, CoT is more like an explanation to the models' hidden belief, instead of reasoning.

Type III. Full Connection is the closest to the nature of autoregressive statistical models, where Y depends on the Z and X in a left-to-right manner, capturing the causal mask property but representing statistical correlation rather than genuine causal relationships.

Type IV. Isolation represents that the answer is influenced solely by the question itself, where neither Z nor X determines the answer. Under this type of SCM, the model resembles memorizing the answer rather than genuinely performing reasoning.

D. Connection between SCM, Latent Behavior, Consistency, and Faithfulness

Causal chain (type I) presents the ideal causal structure, where instruction (\rightarrow thinking) \rightarrow CoT \rightarrow answer. Within this framework, Z is precluded from affecting Y, while both the presence and absence of T's effect on Y are reasonable. When a model works under causal chain, it does real reasoning, producing faithful and consistent output. In contrast, common cause (type II) represents an undesirable underlying structure, where the CoT is not a faithful reflection of models' answerderivation process but a post-hoc explanation conditioned on the latent belief; on the other hand, the answer and the CoT are not necessarily consistent. Generally speaking, all language

Model	Type	G-44*		M	ath		Logic			
Model	Type	Setting	Add.	Mult.	Math500	GSM8K	ProofWriter	FOLIO	LogiQA	
		default	0.998	0.816	-	0.946	0.708	0.686	0.688	
		random CoT	0.994	0.040	-	0.932	0.626	0.637	0.683	
GPT-4	LLM	R-ATE: $X \rightarrow Y$	$0.60\%_{F}$	$95.3\%_{T}$	-	$1.7\%_{F}$	$11.6\%_{T}$	$7.1\%_{F}$	$0.3\%_{F}$	
GP1-4	LLIVI	Instruction bias	0.826	0.848	-	0.946	0.703	0.671	0.675	
		R-ATE: Z→Y	$17.4\%_{T}$	$1.2\%_{F}$	-	$0.2\%_{F}$	$0.7\%_{F}$	$2.2\%_F$	$1.5\%_{F}$	
		SCM Type	II	I	-	IV	I	IV	IV	
		default	0.996	0.99	0.962	-	0.893	0.799	0.807	
		random Thinking	0.994	0.988	0.956	-	0.887	0.784	0.806	
		R-ATE: $T \rightarrow Y$	$0.2\%_{F}$	$0.2\%_{F}$	$0.6\%_{F}$	-	$0.6\%_{F}$	$1.9\%_{F}$	$0.1\%_{F}$	
		random CoT	0.278	0.652	0.858	-	0.793	0.637	0.740	
Deepseek-R1	LRM	R-ATE: $X \rightarrow Y$	$72.1\%_{T}$	$34.1\%_{T}$	$10.8\%_{T}$	-	$11.2\%_{T}$	$20.3\%_{T}$	$8.3\%_{T}$	
·		Instruction bias	0.978	0.982	0.960	-	0.887	0.794	0.812	
		R-ATE: Z→Y	$1.8\%_{T}$	$0.8\%_{F}$	$0.2\%_{F}$	-	$0.7\%_{F}$	$0.6\%_{F}$	0_F	
		SCM Type	III	I	I	-	I	I	I	
				TABLE	ΕI					

Identification of causal structures in tasks running on GPT-4 and DeepSeek-R1, as representatives of LLMs and LRMs, respectively. We present task accuracy and R-ATE, where the subscript 'T/F' denotes statistical significance with p < 0.01 in McNemar's test. "Default" refers to the default CoT generated by the LLM; "random CoT" and "random thinking" indicate that we intervened with random content in parts of the CoT or thinking processes; "instruction bias" represents the introduction of bias into the instruction. Finally, we determine SCM types based on statistical significance.

models are fully connected (type III) to some extent due to the auto-regressive design. Thus, the real latent behavior of a LLM is most likely a mixture of reasoning and explaining. However, we can differentiate them using statistical significance as an indicator, obtaining the most likely underlying causal structures of an LLM on different tasks. Additionally, isolation (type IV) represents a prototypical erroneous SCM, wherein the model's response is determined exclusively by the input question, remaining uninfluenced by any other variables within the causal framework. Under this structure, the model relies on memorized question-answer mappings rather than engaging in genuine reasoning processes.

IV. EXPERIMENTS

We conduct extensive experiments with commercial and open-source LLMs and LRMs, testing the causal structures of the reasoning processes on different math and logical reasoning tasks.

A. Experimental Settings

- 1) Models: We test both non-reasoning models and reasoning models. As for non-reasoning models, we choose Chat-GPT [58], GPT-4 [59], Llama2-7B/70B-Chat [60] and Qwen-2.5-32B-Instruct [61]. As for reasoning models, we select Deepseek-R1-0528 [4], QwQ-32B [62], DeepSeek-R1-Distill-Qwen-32/7B [4] and DeepSeek-R1-Distill-Llama-8B [4]. For Deepseek-R1, GPT3.5 and GPT4, we utilize the official API to conduct experiments; for other models, we deploy them using Vllm [63].
- 2) Datasets: Our evaluation is conducted on mathematical reasoning and logical reasoning datasets. All selected datasets feature unambiguous ground-truth answers and necessitate chain-of-thought reasoning processes to reach the correct answers. The mathematical reasoning datasets consist of 3-digit multiplication, 9-digit addition, GSM8k [64] and MATH500 [65]. The former two datasets, which were created in our previous work [12], evaluate fundamental arithmetic

computation abilities, while MATH500 and GSM8k assess mathematical problem-solving skills through word problems. Each dataset above contains 500 samples. The logical reasoning datasets include ProofWriter [66], FOLIO [67], and LOGIQA [68]. ProofWriter and FOLIO are both multiple-choice datasets for evaluating deductive reasoning. We randomly draw 600 samples from the 5-hop reasoning development set of ProofWriter. As for FOLIO, we utilize all 204 instances from its development set. LOGIQA is a multiple-choice dataset designed to assess logical reasoning abilities in reading comprehension tasks. We extract 600 instances randomly from the LOGIQA2.0 test set.

3) Inference Settings: For experimental deployment, we conduct experiments with Deepseek-R1 via the official API, while other models are deployed using Vllm [63]. We use a temperature setting of 0 for all experimental conditions unless specific parameters are recommended in the model's technical report. We set the maximum output length to 24,000 tokens, with any inference exceeding this limit being truncated. The details of the prompts can be found in our released code.

B. Deriving SCM from Treatment Experiments

SCMs comprise nodes and edges. Nodes represent the variables in our framework, specifically Z, X, Y, and T, where T appears exclusively in the LRM setting. Edges represent the causal relationships between variables, with $X \rightarrow Y$ indicating that variable X causally determines Y. Within each model category (LLM or LRM), the node set remains fixed, while different edge configurations define distinct SCM types. Edge presence is determined by intervention experiments, where a significant R-ATE indicates the existence of an edge between two nodes. After conducting treatment experiments on all edges under consideration to determine their existence, we obtain the SCM of the model on a specific task. As Table I shows, GPT4 represents the 3-variable LLM with variables Z, X, and Y, while Deepseek-R1 represents the 4-variable LLM with variables Z, T, X, and Y.

Type	Model	Method	SCM-I (the ideal)	SCM-II	SCM-III	SCM-IV	SCM(I) %
	ChatGPT	RLHF	GSM8K FOLIO	Add.	Mult. ProofW. LOGIQA		33%
	GPT-4	RLHF	Mult. ProofW.	Add.		GSM8K FOLIO LOGIQA	33%
LLM	Llama-70B-Chat	RLHF	Add. ProofW.			Mult. GSM8K FOLIO LOGIQA	33%
	Llama-7B-Chat	RLHF		Add. ProofW. FOLIO LOGIQA	GSM8K	Mult.	0
	Qwen2.5-32B-Instruct	Instr.	Mult. ProofW. FOLIO		Add. MATH500. LOGIQA		50%
		Avg. Ratio	30%	20%	23%	27%	30%
1			Add.(a)		MATH500(a)		
	R1-Distill-Qwen-7B	Distill	Mult.(b)		ProofW.(a)		50%
			FOLIO(b) Add.(a)		LOGIQA(c) Mult.(a)		
	R1-Distill-Qwen-32B	Distill	MATH500(b)		ProofW.(a)		50%
			FOLIO(b) Mult.(a)		LOGIQA(a)		
	R1-Distill-Llama-8B	Distill	MATH500(b)	LOGIOA(a)	Add.(a)		50%
LRM			FOLIO(a) Mult.(a)		ProofW.(b)		
LIXIVI	D 1.D1		MATH500(a)		A 11 (1)		92.0
	Deepseek-R1	misc.	ProofW.(a) FOLIO(a)		Add.(b)		83%
			LOGIQA(a)				
			Add.(b)				
			Mult.(a)				
	QwQ-32B	RLVR	MATH500(b)			LOGIQA(a)	83%
			ProofW.(b) FOLIO(b)				
	<u> </u> 	Avg. Ratio	63%	3%	30%	3%	63%

TABLE II

COMPARISON OF SCM TYPE DISTRIBUTIONS BETWEEN LLMS AND LRMS. WE ASSESS THE SCM TYPES OF DIFFERENT DATASETS ACROSS MODELS AND CALCULATE THE DISTRIBUTIONS FOR LLMS AND LRMS, RESPECTIVELY. THE FIRST THREE COLUMNS PRESENT MODEL TYPE, MODEL NAME, AND PRIMARY TRAINING METHOD. IN COLUMNS 4-7, WE REPORT THE SCM TYPES OF THE TESTED MATHEMATICAL AND LOGICAL DATASETS, WHERE SCM-I REPRESENTS THE IDEAL CASE. IN THE FINAL COLUMN, WE REPORT THE PROPORTION OF DATASETS CLASSIFIED AS SCM-I FOR EACH MODEL, WITH HIGHER VALUES INDICATING STRONGER CAUSAL REASONING CAPABILITIES. ADDITIONALLY, ROWS 6 AND 12 PRESENT THE AVERAGE PROPORTIONS OF DIFFERENT SCM TYPES FOR LLMS AND LRMS RESPECTIVELY, ENABLING FINE-GRAINED ANALYSIS OF REASONING BEHAVIORS IN THESE TWO MODEL CATEGORIES.

We take Deepseek-R1 on the Mult. dataset as an example for showing how the SCMs are obtained. For the DeepSeek-R1 default condition, the model's accuracy on Mult. is 0.99, serving as the pre-intervention baseline. In the random Thinking, random CoT and Instruction bias conditions, we calculate the post-intervention accuracies following interventions on the Thinking, CoT, and Instruction variables, obtaining 0.988, 0.652, and 0.982, respectively. Using Equation 3, we compute the R-ATE values for interventions on these three variables, obtaining 0.2%, 34.1%, and 0.8%, respectively. According to McNemar's test [57], only the CoT intervention produces a statistically significant change, while interventions on Thinking and Instruction (Z) are not significant. Consequently, the SCM includes an edge from CoT(X) to Answer (Y), but excludes edges from Thinking (T) to Answer (Y) and Instruction (Z)to Answer. As shown in Figure 2, this SCM corresponds to Type-I SCM category (a).

C. Results of Causal Structures

We separately calculate the SCM distributions for LLMs and LRMs in Table II. The findings are as follows. First, LLM models belonging to the ideal SCM-I average only 30%, while the corresponding value for LRMs is 63%. Notably, even the weakest model among LRMs possesses 3 type-I SCMs, which equals the maximum number of type-I SCMs found in any LLM model. This demonstrates that LRM models exhibit significantly stronger causality than LLMs.

Among LLMs, SCM-I accounts for the largest proportion, demonstrating a certain degree of causality. However, the SCM types are also relatively evenly distributed. Specifically, 20% belong to SCM-II explaining and 27% belong to SCM-IV memorizing. In both of these SCM types (combined 47%), CoT cannot significantly influence the results, indicating that LLM's CoT is largely unfaithful. The remaining 23% fall under SCM-III, which represents a mixture of reasoning

and common case. SCM-II and SCM-III (combined 43%), results are affected by ineffective conditions in the instruction, reflecting the inherent nature of autoregressive models.

For LRMs, the majority (63%) of SCMs are ideal SCM-I, with the remaining most frequent type being SCM-III (30%). This indicates that LRM's CoT has a very stable and significant influence on answers (combined 93%). The remaining issues are concentrated on the influence of instructionirrelevant conditions on results in SCM-III. Additionally, 3% each belong to SCM-II and SCM-IV (combined 6%), reflecting unfaithful CoT in a small number of cases, but this issue is significantly less prevalent than in LLMs. Notably, the vast majority of non-ideal cases among the tested LRMs occur in distilled LRMs, whereas OwO-32B which relies almost entirely on RLVR, and DeepSeek-R1 which combines distillation with RLVR, exhibit almost no non-ideal SCMs, further demonstrating the positive impact of RLVR on causality. Moreover, a direct comparison among QwQ-32B (RLVR with thinking), R1-DeepSeek-Qwen-32B (SFT with thinking), and Owen2.5-32B-Instruct (SFT without thinking) suggests that the presence or absence of Thinking (T) has limited effect on causality, whereas the training paradigm, particularly RLVR, plays the dominant role. Inspired by these observations, we further investigate the impact of different learning paradigms in Section V.

We further examine the four-variable SCM minor types for LRMs, where the distinction from three-variable SCMs lies in the decomposition of LRM's reasoning process into two separate variables: T and X. Notably, the thinking variable significantly influences the answer in Type-I (b) and Type-III (a) and (c) configurations. The proportion of datasets belonging to these SCM categories is 67% for both Qwenbased distilled models and 67% for QwQ-32B, while R1-Distill-Llama-8B and DeepSeek-R1 show lower proportions at 33% each. This indicates that the latter two LRMs exhibit more linear dependence on the preceding variable X for answer generation, whereas the former three LRMs demonstrate simultaneous dependence on both T and X variables.

D. Summary of Findings

The CoT reasoning for LLMs is not causal but rather statistical. It is not only susceptible to interference from extraneous information in instructions, but also exhibit a substantial number of SCMs where CoT is unreliable. Our findings explain experimental observations from previous work showing that incorrect CoT can lead to correct conclusions, and vice versa [9], [13], [69]. LRM models exhibit significantly superior causality compared to LLMs. The credibility of LRM's CoT is substantially higher than that of LLMs, though it is still somewhat influenced by the inherent autoregressive nature, causing answers to be disturbed by irrelevant information in instructions. Among LRMs, those trained with RLVR significantly outperform distilled LRMs, demonstrating the positive impact of RLVR on causality.

V. LEARNING TECHNIQUES SHAPE CAUSAL STRUCTURES

Building on the previous section, we investigate how different learning paradigms shape the causal structures in language

models. We systematically evaluate in-context learning (ICL), supervised fine-tuning (SFT), distillation, reinforcement learning (RL), and their combinations (SFT+RL and Distill+RL), assessing both causal structures and robustness. This allows us to quantify the contribution of each paradigm to causal alignment in reasoning models.

A. Impact of In-Context Learning

In-context learning (ICL) is widely adopted to elicit desired behaviors in LLMs, and is commonly used to trigger chainof-thought reasoning to mimic human step-by-step inference. However, ICL techniques such as few-shot prompting are generally not recommended for LRMs [4], therefore we only evaluate the effectiveness of ICL on LLMs in our experiments.

Settings. We evaluate ICL using GPT-3.5-Turbo as the base model. Following the same set of tasks described in Section IV-A, we prepend k randomly sampled demonstrations to the prompt, with k varied across conditions. All demonstrations are drawn from the development set without leakage into evaluation examples. For each condition, we conduct intervention experiments and compute SCM structures, R-ATEs, and task accuracies.

Results and Analysis. As shown in Table III, compared to the zero-shot setting, ICL improves both task accuracy and causal alignment slightly. Specifically, ICL reduces the absolute R-ATE of the Instruction \rightarrow Answer edge while enhancing that of $CoT \rightarrow Answer$, leading to more coherent causal structures. This suggests that demonstrations help suppress spurious correlations and strengthen reasoning grounded in the CoT.

B. Impact of SFT and RLHF

Supervised fine-tuning (SFT) enables LLMs to better follow human instructions, while reinforcement learning from human feedback (RLHF) further aligns model behavior with human preferences. However, recent studies have also shown that both SFT and RLHF may increase hallucinations and unfaithfulness [70], [71]. We therefore hypothesize that these paradigms may influence the underlying causal structures of reasoning.

Settings. To validate this hypothesis, we analyze three models from the Mistral family: Mistral-7B-Base, Mistral-7B-SFT, and Mistral-7B-DPO [72]. Since the base model cannot reliably follow instructions, we elicit its question-answering behavior using ICL with four demonstrations. For consistency, the same demonstrations are used for the SFT and DPO models. We then perform intervention experiments and compute SCM structures, R-ATEs, and task accuracies, following the same procedure as in Section IV-A.

Results and Analysis. As shown in Table IV, SFT generally weakens causal alignment, though individual models may exhibit specific variations. It reduces the average R-ATE on the $CoT \rightarrow Answer$ edge while increasing that on the Instruction \rightarrow Answer edge, suggesting that SFT introduces spurious dependencies between instructions and answers, which in turn may cause hallucinations. In contrast, RLHF (via DPO) mitigates spurious correlations by reducing the strength of the Instruction \rightarrow Answer edge and lowering the average R-ATE

ICL	Metric	Add.	Mult.	GSM.	Pro.	FOL.	LQA.	AVG	SCM-I	Task Acc
0-shot		$ \begin{array}{ c c } \hline 2.3\%_F \\ 70.8\%_T \\ \hline \text{II} \\ \hline \end{array} $	$100.0\%_{T} \ 9.3\%_{T} \ \mathrm{III}$	$\begin{array}{c} 97.6\%_T \\ 0_F \\ \mathrm{I} \end{array}$	$17.9\%_{T} \ 3.3\%_{T} \ \mathrm{III}$	$15.8\%_{T} \ 4.3\%_{F} \ { m I}$	$8.3\%_{T}$ $9.8\%_{T}$ III	40.3% 16.2%	2	0.572
2-shot	$ \begin{array}{c c} R\text{-ATE: } X \rightarrow Y \uparrow \\ R\text{-ATE: } Z \rightarrow Y \downarrow \\ SCM \end{array} $	$\begin{array}{ c c }\hline 0_F \\ 17.6\%_T \\ \text{II} \end{array}$	$100.0\%_T \\ 0_F \\ \mathrm{I}$	$99.2\%_{T}$ 0_{F} I	$45.6\%_{T}$ $19.9\%_{T}$ III	$46.6\%_{T}$ $15.5\%_{T}$ III	$8.4\%_{T} \\ 1.9\%_{F} \\ \mathrm{I}$	50.0% 9.2%	3	0.598
4-shot	$ \begin{array}{c c} R\text{-ATE: } X \rightarrow Y \uparrow \\ R\text{-ATE: } Z \rightarrow Y \downarrow \\ SCM \end{array} $	$\begin{array}{c c} 0\%_F \\ 0\%_F \\ \text{IV} \end{array}$	$100.0\%_T \\ 0\%_F \\ \mathrm{I}$	$99.5\%_T \\ 0\%_F \\ \mathrm{I}$	$40.9\%_{T}$ $13.2\%_{T}$ III	$39.0\%_{T} \ 20.4\%_{T} \ \mathrm{III}$	$\begin{array}{c} 4.4\%_F \\ 0\%_F \\ \text{IV} \end{array}$	47.3% 5.7%	2	0.580
8-shot		$\begin{array}{ c c }\hline 0\%_F\\0\%_F\\\text{IV}\end{array}$	$99.7\%_{T} \\ 0\%_{F} \\ \mathrm{I}$	$99.2\%_T \\ 0\%_F \\ \mathrm{I}$	$47.1\%_{T} \ 32.9\%_{T} \ \mathrm{III}$	$30.9\%_{T} \ 25.2\%_{T} \ \mathrm{III}$	$6.8\%_T \\ 1.6\%_F \\ \mathrm{I}$	47.3% 10.0% -	3	0.592
				TAB	LE III					

The impact of ICL on causal relationships tested on GPT-3.5-Turbo, where the best |R-ATE| and task accuracy are marked in bold. The 'T/F' indicates the statistical significance of the causal relation.

Model	Metric	Add.	Mult.	GSM.	Pro.	FOL.	LQA.	AVG	SCM-I	Task Acc
Base		$\begin{array}{ c c }\hline 0\%_F\\ 100\%_F\\ \text{IV}\end{array}$	$100.0\%_{T} \ 6.8\%_{F} \ { m I}$	$98.2\%_{T} \ 2.3\%_{F} \ { m I}$	$32.0\%_{T} \ 87.8\%_{T} \ \mathrm{III}$	$0\%_F \\ 24.4\%_F \\ \mathrm{IV}$	$9.0\%_T \\ 0\%_F \\ \mathrm{I}$	39.9% 36.9%	3	0.313
SFT		$\begin{array}{ c c } & 0_F \\ 100\%_T \\ & \text{II} \end{array}$	$92.3\%_{T} \\ 57.7_{T} \\ \text{III}$	$97.1\%_T \\ 9.5_T \\ \text{III}$	$\begin{array}{c} 0\%_F \\ 12.0\%_T \\ \text{II} \end{array}$	$0\%_F\\34.6\%_T\\\text{II}$	$5\%_F \\ 22.8\%_T \\ \mathrm{II}$	32.4% 39.4% -	0	0.300
DPO	$ \begin{array}{c c} R\text{-ATE: } X {\rightarrow} Y \uparrow \\ R\text{-ATE: } Z {\rightarrow} Y \downarrow \\ SCM \end{array} $	$\begin{array}{c c} 0\%_F \\ 0\%_F \\ \text{IV} \end{array}$	$60.0\%_F \\ 60.0\%_F \\ \mathrm{IV}$	$96.3\%_{T} \ 8.6\%_{F} \ { m I}$	$0\%_F \\ 11.3\%_F \\ \mathrm{IV}$	$12.3\%_F \\ 8.6\%_F \\ \mathrm{IV}$	$12.0\%_{T}$ $29.9\%_{T}$ III	30.1% 19.7 %	1	0.275
				TAI	BLE IV					

The impact of SFT/RLHF ON CAUSAL RELATIONSHIPS BASED ON Mistral-7B, WHERE SFT PRIMARILY ENHANCES THE CAUSAL CONNECTION BETWEEN THE INSTRUCTION AND ANSWER, AND DPO DIMINISHES THIS RELATIONSHIP.

$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	Model	Metric	Add.	Mult.	MATH500	Pro.	FOL.	LQA.	AVG	SCM-I	Task Acc
R1-Distill-Qwen-32B R-ATE: $Z \rightarrow Y \downarrow$ $2.2\%_F$ $2.3\%_T$ 1.4_F $6.1\%_T$ $0.6\%_F$ $2.6\%_T$ 2.5% 3 0.84	Qwen2.5-32B-Instr.	R-ATE: Z→Y ↓	$19.7\%_{T}$		$8.6\%_{T}$			$2.2\%_{T}^{2}$	5.3%	3	0.76
	0.84										

Comparison of Distill and SFT on Causal relationships based on Qwen-32B-Instruct and R1-Distill-Qwen-32B, where compared to Distill and Instruction-tuning, there are no significant causal changes.

from 36.9% to 19.7%, consistent with the human preference to separate answers from irrelevant instructions [20]. However, DPO also weakens the $CoT \rightarrow Answer$ link (average R-ATE reduced to 0.138), indicating a trade-off where RLHF suppresses spurious patterns but may also diminish genuine causal connections.

C. Impact of Distillation

Distillation has become a dominant paradigm in recent LRMs, where smaller models learn the reasoning traces from stronger teachers. This approach is proved to yield stronger performance than RLVR training [4]. Here, we analyze whether such gains also improve causal reasoning. Since distillation is technically a form of SFT, we compare it against instruction-tuning, another commonly used SFT approach.

Settings. We analyze three LRMs distilled with DeepSeek-R1 data: R1-Distill-Qwen-7B, R1-Distill-Qwen-32B, and R1-

Distill-Llama-8B, and compare them against the Instructiontuned model *Qwen2.5-32B-Instruct*. Notably, both *R1-Distill-Qwen-32B* and *Qwen2.5-32B-Instruct* are derived from the same base model Qwen2.5-32B.

Results and Analysis. As shown in Table I, the three distilled LRMs exhibit same proportions of Ideal SCMs, all achieving 50%, indicating consistent performance across distilled LRMs. We further compare Qwen2.5-32B-Instruct and R1-Distill-Qwen-32B, with detailed results presented in Table V. The results show that distillation significantly improves average task accuracy compared to instruction-tuning, which aligns with findings from other studies [4]. However, both models have an identical number (3) of type-I SCMs, and the R-ATE changes for X→Y and Z→Y edges are minimal. This suggests that while distilled LRMs enhance task performance, they do not improve causality, showing no distinction from conventional instruction-tuning at the causal level.

Original Question	Modified Question							
If $f(x) = \frac{3x-2}{x-2}$, what is the value of $f(-2) + f(-1) + f(0)$? Express	If $f(x) = \frac{3x-2}{x-2}$, and given that $a = 7$ and $b = -5$, what is the							
your answer as a common fraction.	value of $f(-2) + f(-1) + f(0)$? Express your answer as a common							
	fraction.							
TABLE VI								

COMPARISON OF ORIGINAL AND MODIFIED QUESTIONS. THE BOLD PORTION "AND GIVEN THAT A=7 AND B=-5" REPRESENTS EXTRANEOUS CONDITIONS IRRELEVANT TO PROBLEM SOLVING, ADDED IN MATH500-NOOP TO TEST WHETHER MODELS RELY ON SUPERFICIAL FEATURES.

	Math500 Acc: High	Math500 Acc: Low						
Math500-Noop Acc: High	Good fitting, few spurious features, strong generalization, strong causality	Rare state, not discussed						
Math500 Noon Aga Law	Good fitting, many spurious features,	Underfitting, few spurious features						
Math500-Noop Acc: Low	poor generalization, weak causality	but also few genuine features, weak causality						
TABLE VII								

ANALYSIS OF MODEL PERFORMANCE ACROSS DIFFERENT FEATURE COMBINATIONS

Group	Base Model	Comparison Setting	Source	Algorithm	Dataset	Purpose
Base-RLVR	Qwen2.5-3B-Base	Qwen2.5-3B-Base	Open source	-	-	Base vs Base+RLVR
		Qwen2.5-3B-Base-GRPO-600/1000/2000 steps	Self-trained	GRPO	DeepScaleR	
SFT-RLVR	Qwen2.5-3B-	Qwen2.5-3B-Instruct	Open source	Instr.	-	SFT vs SFT+RL
SI I REVI	Instruct	Qwen2.5-3B-Instruct-GRPO-200/1000/2000 steps	Self-trained	GRPO	DeepScaleR	577 (0 577) (12
Base-	Owen2.5-3B-Base	Qwen2.5-3B-Base-Distill-Openr1	Self-trained	Distill	OpenR1	Base+RL vs Base+Distill
RLVR/Distill	Q Well 210 0 B Base	Qwen2.5-3B-Base-GRPO-Openr1	Self-trained	GRPO	OpenR1	Base File vs Base Filsani
Distill-	Qwen2.5-Math-	Deepseek-Distill-Qwen-1.5B	Open source	Distill	-	Distill vs Distill+RL
RLVR	1.5B	DeepscaleR-1.5B-Preview	Open source	Distill+GRPO	-	Distill 15 Distill IRE

TABLE VIII

WE UTILIZE FOUR MODEL GROUPS TO ELUCIDATE WHY RLVR IMPROVES CAUSAL REASONING IN MODELS. DETAILS REGARDING THE SOURCE, ALGORITHM, DATASET, AND EXPERIMENTAL OBJECTIVES FOR EACH GROUP ARE PROVIDED IN THIS TABLE.

D. Impact of RLVR

Reinforcement learning from verifiable rewards (RLVR) has recently emerged as a key optimization paradigm in LRMs, driving substantial gains in reasoning performance. Unlike RLHF, which relies on subjective human preference signals, RLVR leverages automatically verifiable outcomes, the correctness of the final answer, as reward signals. By directly optimizing models with respect to correctness, RLVR not only improves task accuracy but also refines the reasoning process, making it more consistent with causal reasoning structures. In this subsection, we analyze how RLVR shapes causality under the following SFT+GRPO settings.

Settings. We apply RLVR using the GRPO algorithm on Qwen2.5-3B-Instruction, following the outcome reward method of DeepScaleR [73]. Given the relatively small scale of the model, we evaluate it on 6-digit Addition, 2-digit Multiplication, Math500, ProofWriter, FOLIO and LOGIQA. The DeepScaleR dataset consists of math word problems similar to MATH500, making MATH500 an in-domain task. Addition and Multiplication can be considered out-of-distribution (OOD) arithmetic problems, while ProofWriter, FOLIO and LOGIQA represent OOD logical reasoning problems.

Results and Analysis. According to Table IX, GRPO improves task accuracy from 0.59 to 0.62. Simultaneously, the number of Type-I SCMs increases from 1 to 4. Among all

the methods discussed above, GRPO achieves the largest gain in causality. This causal enhancement can also be observed in Table II. QwQ-32B-Instruct achieves Type-I SCMs in 5 tasks, substantially outperforming R1-Distill-Qwen-32B and Qwen2.5-32B, which employ distillation or instruction-tuning on the same base model.

Beyond the in-domain MATH500, *Qwen2.5-3B-Instruct-GRPO* demonstrates substantial out-of-distribution gains on logical reasoning benchmarks, despite being trained exclusively on mathematical data. As shown in Table IX, the baseline prior to GRPO achieves a Type-I SCM only on FOLIO, whereas GRPO upgrades *all three* OOD logic tasks (ProofWriter, FOLIO, LOGIQA) to Type-I. This improvement is primarily driven by suppressing the spurious $Z \rightarrow Y$ edge—from 11.2%, 3.8%, and 3.1% to 1.5%, 0%, and 1.9% on Pro./FOL./LQA., respectively—while maintaining strong $X \rightarrow Y$ links across tasks. These results provide compelling evidence that RLVR learns transferable causal patterns rather than memorizing task-specific correlations [74].

VI. HOW DOES RLVR ENHANCE CAUSALITY

The analyses in the previous section demonstrate that RLVR enhances causal reasoning structures significantly. Building on these observations, we now aim to explain how RLVR leads to ideal causality. Intuitively, causality is closely tied to

		Add.	Mult.	MATH500	Pro.	FOL.	LQA.	AVG	SCM-I	Task Acc
1	R-ATE: $X \rightarrow Y \uparrow$ R-ATE: $Z \rightarrow Y \downarrow$ SCM	$69.3\%_{T} \\ 7.2\%_{T} \\ \text{III}$	$72.3\%_{T} \\ 6.3\%_{T} \\ \text{III}$	$90.2\%_{T}$ $13.1\%_{T}$ III	$49.9\%_T \\ 11.2\%_T \\ \text{III}$	$58.1\%_{T} \\ 3.8\%_{F} \\ \mathrm{I}$	$27.3\%_T \\ 3.1\%_T \\ \mathrm{III}$	61.2 % 7.5%	1	0.59
	R-ATE: $X \rightarrow Y \uparrow$ R-ATE: $Z \rightarrow Y \downarrow$ SCM	$93.8_T \\ 20.4\%_T \\ \text{III}$	$72.4\%_{T}$ $9.0\%_{T}$ III	$72.9\%_{T} \\ 0_{F} \\ \mathrm{I}$	$39.2\%_{T}$ $1.5\%_{F}$ I	$46.0\%_T \\ 0\%_F \\ \mathrm{I}$	$27.7\%_{T}$ $1.9\%_{F}$ I	58.7% 5.5 %	4	0.62

The impact of RLVR on Causal relationships based on Qwen2.5-3B-Instruct, where GRPO significantly increased the number of good SCMs and improved task performance.

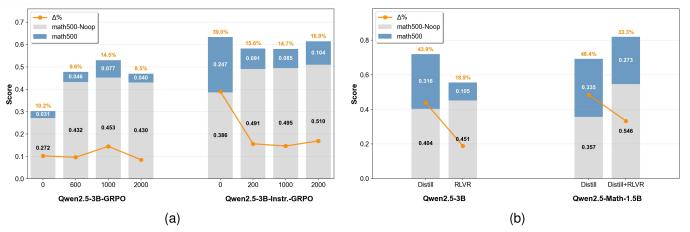


Fig. 3. We evaluate the performance of four groups of models: Base-RL, SFT-RL, Base-RL/Distill, and Distill-RL, on Math500 and Math500-Noop. In the Base-RLVR group, 0, 600, 1000, and 2000 represent Qwen2.5-3B and checkpoints of Qwen2.5-3B after 600, 1000, and 2000 GRPO steps, respectively. In the SFT-RLVR group, 0, 200, 1000, and 2000 represent Qwen2.5-3B-Instruct and checkpoints of Qwen2.5-3B-Instruct after 200, 1000, and 2000 GRPO steps, respectively. In the Base-RLVR/Distill group, "Distill" represents the Qwen2.5-3B model distilled with OpenR1-Math-220k data, while "RLVR" represents the model obtained through GRPO using the same base architecture. In the Distill-RLVR group, "Distill" refers to the open-source Deepseek-Distill-Qwen-1.5B model, which is trained from Qwen2.5-Math-1.5B using distillation, while "Distill+RLVR" represents the open-source DeepsealeR-1.5B-Preview model, which applies GRPO on top of Deepseek-Distill-Qwen-1.5B. The bar chart represents the accuracy of Math500 and Math500-Noop, while the line chart depicts the relative difference between Math500 and Math500-Noop to measure the model's generalization capability and the extent to which genuine features are learned; the relative gap in accuracy between Math500 and Math500-Noop is employed to quantify the degree of spurious feature learning. (a) Contains Base-RLVR and SFT-RLVR, which reflect the changes in genuine and spurious features when applying GRPO starting from Base and SFT models, respectively. (b) Contains Base-RLVR/Distill and Distill-RLVR. The former reflects the changes in genuine and spurious features between RLVR and Distill under identical training data, while the latter reflects the scenario of applying RLVR on top of the Distill model.

the contrast between genuine and spurious features: genuine features are causally related to the true label, while spurious features lack causal connection to the label but are coincidentally correlated with it in the data [17], [75]. Models exhibit better causal alignment when reasoning relies more on genuine features and less on spurious ones [17]. To validate this hypothesis, we design three sets of comparative experiments.

A. Measuring Spurious Feature Reliance

To quantify the extent to which models exploit spurious correlations, we construct a variant of Math500 following the GSM-Noop methodology [76]. The dataset, termed *Math500-Noop*, augments each problem with 1–2 irrelevant numerical conditions generated by GPT-4.1-mini. An example is shown in Table VI. Since the added conditions are causally unrelated to the solution, a model that relies on pattern matching or shortcut heuristics may be misled, whereas a model that performs genuine reasoning should remain robust. We then define the metric:

$$\Delta\% = \frac{Acc(\text{Math500}) - Acc(\text{Math500-Noop})}{Acc(\text{Math500})} \tag{7}$$

with larger $\Delta\%$ indicating stronger spurious reliance. Table VII summarizes how combinations of Math500 and Math500-Noop accuracies map to causality levels: strong causality requires both high fitting (high Acc on Math500 and Math500-Noop) and low spurious reliance (small $\Delta\%$), while other cases correspond to weak causality.

B. Correlation Between Causality and Spurious Features

We now investigate whether the CoT causality and the proportion of the spurious features are quantitatively correlated during RLVR training. Under the assumption that stronger causality arises from reduced reliance on spurious features, metrics of causal alignment and spurious feature dependence should exhibit consistent correlation across training checkpoints. To validate this, we explicitly define how to measure spurious feature reliance and causal alignment, and then compute their correlation over multiple RLVR training trajectories.

Settings. We analyze two model groups: Base-RL, SFT-RL. For each group, we collect checkpoints along the RLVR training trajectory and compute two complementary metrics. The first is the number of type-I SCMs, derived from the

Model	Metric	indomain	0	OD-Arithmetic			OOD-	logic		Cood SCM(I)
Model	Metric	Math500	Add	Mult	avg	ProofWriter	FOLIO	LOGIQA	avg	Good SCM(I)
	CoT→Ans: R-ATE↑	$79.39\%_{T}$	$5.06\%_{F}$	$47.95\%_{T}$	26.51%	$39.31\%_{T}$	$63.06\%_{T}$	$12.42\%_{T}$	38.26%	-
Owen2.5-3B-Base	Instr→Ans: R-ATE↓	$17.94\%_{T}$	$94.90\%_{T}$	$87.67\%_{T}$	91.29%	$46.82\%_{T}$	$18.92\%_{T}$	$69.57\%_{T}$	45.10%	-
Qweii2.3-3B-Base	SCM	III	II	III	-	III	III	III	-	0
	CoT→Ans: R-ATE↑	$79.52\%_{T}$	$98.20\%_{T}$	$96.65\%_{T}$	97.43%	$48.21\%_{T}$	$62.04\%_{T}$	$22.06\%_{T}$	44.10%	-
(GRPO step 600)	Instr→Ans: R-ATE↓	$5.85\%_{T}$	$44.01\%_{T}$	$10.05\%_{T}$	27.03%	$6.92\%_{T}$	$1.76\%_{F}$	$22.06\%_{T}$	10.25%	-
(GKPO step 600)	SCM	III	III	III	-	III	I	III	-	1
	CoT→Ans: R-ATE↑	$74.51\%_{T}$	$100.00\%_{T}$	$84.87\%_{T}$	92.44%	$39.07\%_{T}$	$51.49\%_{T}$	$27.00\%_{T}$	39.19%	-
(GRPO step 1000)	Instr→Ans: R-ATE↓	$6.77\%_{T}$	$17.84\%_{T}$	$3.59\%_{T}$	10.72%	$11.55\%_{T}$	$10.69\%_{T}$	$16.25\%_{T}$	12.83%	-
	SCM	III	III	III	-	III	III	III	-	0
	CoT→Ans: R-ATE↑	$87.73\%_{T}$	$97.79\%_{T}$	$80.36\%_{T}$	89.08%	$58.32\%_{T}$	$53.90\%_{T}$	$11.34\%_{T}$	41.19%	-
(GRPO Step 2000)	Instr→Ans: R-ATE↓	$0.20\%_{F}$	$59.38\%_{T}$	$16.54\%_{T}$	37.96%	$4.93\%_{T}$	$0.89\%_{F}$	$10.42\%_{T}$	5.41%	-
	SCM	I	III	III	-	III	I	III	-	2
	CoT→Ans: R-ATE↑	$90.18\%_{T}$	$69.28\%_{T}$	$72.30\%_{T}$	70.79%	$49.91\%_{T}$	$58.07\%_{T}$	$27.29\%_{T}$	45.09%	-
Owen2.5-3B-Instr	Instr→Ans: R-ATE↓	$13.09\%_{T}$	$7.23\%_{T}$	$6.33\%_{T}$	6.78%	$11.19\%_{T}$	$3.80\%_{F}$	$3.05\%_{T}$	6.01%	-
Qweii2.3-3b-iiisii	SCM	III	III	III	-	III	I	III	-	1
	CoT→Ans: R-ATE↑	$69.57\%_{T}$	$99.78\%_{T}$	$78.90\%_{T}$	89.34%	$15.23\%_{T}$	$54.39\%_{T}$	$0\%_{F}$	23.21%	-
(GRPO Step 200)	Instr→Ans: R-ATE↓	$2.80\%_{F}$	$9.89\%_{T}$	$8.12\%_{T}$	9.01%	$21.87\%_{T}$	$2.49\%_{F}$	$1.29\%_{F}$	8.55%	-
(OKFO Step 200)	SCM	I	III	III	-	III	I	IV	-	2
	CoT→Ans: R-ATE↑	$75.99\%_{T}$	$94.59\%_{T}$	$75.07\%_{T}$	84.83%	$39.18\%_{T}$	$58.54\%_{T}$	$34.95\%_{T}$	44.22%	-
(CDDO Stan 1000)	Instr→Ans: R-ATE↓	$0\%_F$	$58.11\%_{T}$	$8.80\%_{T}$	33.46%	$2.04\%_{F}$	$0.83\%_{F}$	$6.81\%_{T}$	3.23%	-
(GRPO Step 1000)	SCM	I	III	III	-	I	I	III	-	3
	CoT→Ans: R-ATE↑	$72.93\%_{T}$	$93.75\%_{T}$	$72.38\%_{T}$	83.07%	$39.23\%_{T}$	$46.04\%_{T}$	$27.66\%_{T}$	37.64%	-
(CDDO Stan 2000)	Instr→Ans: R-ATE↓	$0\%_{F}$	$20.43\%_{T}$	$9.01\%_{T}$	14.72%	$1.47\%_{F}$	$0\%_F$	$1.87\%_{F}$	1.11%	-
(GRPO Step 2000)	SCM	I	III	III	-	I	I	I	-	4
	•	•	•	TABLE	ΞX	•				

THE NUMBER OF TYPE-I SCMS GENERALLY INCREASES DURING RL TRAINING PROCESS.

intervention-based SCM framework in Section IV-A, which reflects the extent of causal alignment. The second is the relative accuracy difference $\Delta\%$ between Math500 and Math500-Noop, which serves as a proxy for spurious feature reliance. To capture changes induced by RLVR rather than absolute levels, we compute differences for both metrics relative to the initial (pre-RLVR) checkpoint of each trajectory. Finally, we pool all checkpoints from Base-RL and SFT-RL and calculate the Pearson correlation coefficient between these change values. The corresponding models and other details are provided in Table VIII.

Results and Analysis. Across both Base-RL and SFT-RL trajectories, we observe that causal alignment and spurious feature reliance evolve in opposite directions during RLVR training. For the Base-RL group, the number of ideal SCMs fluctuates across checkpoints (Table X), rising from 0 to 1 at step 600, dropping back to 0 at step 1000, and then recovering to 2 at step 2000. Interestingly, these fluctuations are mirrored by changes in spurious feature reliance (Δ %, Fig. 3a): when SCM counts increase, $\Delta\%$ decreases, and vice versa. This pattern suggests that improvements in causal alignment are consistently accompanied by reductions in spurious correlations. The SFT-RL group shows a more stable trajectory: starting with 1 ideal SCM in the baseline, the count increases steadily to 4 after 2000 RLVR steps, while $\Delta\%$ decreases from 39% to around 17%, confirming that RLVR simultaneously strengthens causal structures and suppresses spurious features.

To quantify this relationship, we compute the Pearson correlation between changes in type-I SCM counts and changes in $\Delta\%$, relative to the pre-RLVR baseline. Pooling checkpoints from both Base-RL and SFT-RL, we obtain a coefficient of -0.68(p=0.065), indicating a strong negative correlation that is marginally significant. The consistent trend across trajectories provides robust evidence that improvements in causality under RLVR are tightly linked to the suppression of spurious features.

These findings reinforce our hypothesis that models achieve stronger causal alignment only when reasoning relies more on genuine features and less on superficial correlations. In this sense, Math500-Noop serves as a critical probe—its performance reflects the extent of genuine reasoning, while the gap to Math500 quantifies spurious reliance. Importantly, our results also suggest that neither abundant genuine features alone nor reduced spurious features alone are sufficient; effective causality emerges only when both conditions are simultaneously met.

C. Distillation vs. RLVR

Distillation has recently been advocated such as Deepseek-R1 [4] as a more effective strategy for training smaller reasoning models than RLVR. From a causal perspective, however, it remains unclear whether the gains from distillation reflect genuine reasoning improvements or amplification of spurious correlations. To answer this, we compare distillation and RLVR under controlled conditions and further examine whether RLVR can mitigate the spurious features inherited from distilled checkpoints.

Settings. We consider two setups for comparing distillation and RLVR. First, in a controlled comparison, we train Qwen2.5-3B on OpenR1-Math-220k with both distillation and RLVR, obtaining Qwen2.5-3B-Base-Distill-OpenR1 and Qwen2.5-3B-Base-GRPO-OpenR1. This setting enables a direct evaluation of how the two paradigms affect causal alignment under identical data conditions. Second, to test whether RLVR can refine distilled models, we analyze two open-source checkpoints: Deepseek-Distill-Qwen-1.5B and its RLVR-enhanced counterpart Deepscaler-1.5B-Preview [73]. We present more details of these two groups of models in Table VIII.

To assess reliance on spurious features, we employ the *Math500-Noop* dataset introduced in Section VI-B, and compute the relative accuracy gap between Math500 and Math500-

Noop $(\Delta\%)$ as a proxy for spurious feature dependence. This ensures methodological consistency with the analyses in the previous subsection.

Results and Analysis. In the controlled setting (Base-Distill/RL), we find a clear divergence between the two paradigms. Distillation substantially improves in-distribution accuracy (e.g., Math500), but its performance on Math500-Noop lags behind RLVR despite the latter's lower Math500 accuracy (Figure 3b). This indicates that distillation's gains are largely driven by spurious correlations.

In the Distill-RL group, comparing *Deepseek-Distill-Qwen-1.5B* with *Deepscaler-1.5B-Preview*, we observe that RLVR improves both Math500 and Math500-Noop performance, narrowing the gap between them (Figure 3b). This demonstrates that RLVR can reduce some of the spurious correlations introduced by distillation and enhance causal alignment beyond the distilled baseline.

Overall, while distillation yields strong raw accuracy, it does so at the cost of amplifying spurious features. In contrast, RLVR enhances causal robustness by diminishing reliance on spurious correlations, whether applied directly to base models or used to refine distilled ones.

D. Discussion

Our analyses confirm that robust reasoning requires not only strong task performance but also reliable causal structures. Building on this, we highlight three main insights into the role of RLVR and distillation in shaping causal structures of reasoning models.

First, RLVR improves generalization by systematically reducing reliance on spurious features. This is consistent with prior analyses suggesting that verifiable reward signals encourage models to explore diverse reasoning paths and suppress shortcut heuristics [77], [78]. Importantly, such improvements are not fully reflected in standard task accuracy but become evident on causality-sensitive benchmarks such as Math500-Noop, highlighting the necessity of causal evaluation as a complement to conventional metrics.

Second, distilled models exhibit substantially higher levels of spurious features (Figure 3b). Even after applying RLVR on top of distillation, spurious correlations persist abundantly, suggesting that correlations introduced during distillation are partially irreversible. This limitation indicates that relying solely on accuracy-based evaluation can be misleading: distillation appears effective on in-distribution tasks but weakens robustness in ways that persist even after further RLVR training.

Third, our findings reveal a fundamental trade-off between task accuracy and causal robustness. Distillation consistently achieves higher in-distribution accuracy but amplifies spurious correlations, while RLVR sacrifices some accuracy yet produces stronger causal alignment and robustness under distribution shifts. This raises an open challenge for future work: to design training paradigms that simultaneously achieve strong fitting ability and causal reliability, for example by combining the strengths of distillation and RLVR in hybrid frameworks or by explicitly incorporating causal signals into optimization objectives.

VII. CONCLUSION

We investigate the causal structure of CoT reasoning for LLMs and LRMs. Our findings reveal that LLMs exhibit statistical rather than causal reasoning, whereas LRMs demonstrate superior causality in their reasoning processes. Our experimental evidence shows that RLVR training progressively strengthens genuine causal relationships while mitigating spurious correlations, establishing a clear pathway for developing more reliable reasoning systems. These findings provide crucial insights for causality-driven AI development and highlight reinforcement learning as the preferred training approach when causally robust reasoning is essential for trustworthy AI applications.

REFERENCES

- J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [2] DeepSeek-AI, "Deepseek-v3 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2412.19437
- [3] OpenAI, "Introducing OpenAI o1-preview," https://openai.com/index/ introducing-openai-o1-preview/, 2024.
- [4] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi et al., "Deepseek-r1: Incentivizing reasoning capability in Ilms via reinforcement learning," arXiv preprint arXiv:2501.12948, 2025.
- [5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou et al., "Chain-of-thought prompting elicits reasoning in large language models," Advances in Neural Information Processing Systems, vol. 35, pp. 24824–24837, 2022.
- [6] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22199–22213, 2022.
- [7] L. Wang, W. Xu, Y. Lan, Z. Hu, Y. Lan, R. K.-W. Lee, and E.-P. Lim, "Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models," arXiv preprint arXiv:2305.04091, 2023.
- [8] T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Durmus, E. Hubinger, J. Kernion et al., "Measuring faithfulness in chain-of-thought reasoning," arXiv preprint arXiv:2307.13702, 2023.
- [9] M. Turpin, J. Michael, E. Perez, and S. R. Bowman, "Language models don't always say what they think: Unfaithful explanations in chain-ofthought prompting," arXiv preprint arXiv:2305.04388, 2023.
- [10] O. Bentham, N. Stringham, and A. Marasović, "Chain-of-thought unfaithfulness as disguised accuracy," arXiv preprint arXiv:2402.14897, 2024.
- [11] M. Jin, Q. Yu, H. Zhao, W. Hua, Y. Meng, Y. Zhang, M. Du et al., "The impact of reasoning step length on large language models," arXiv preprint arXiv:2401.04925, 2024.
- [12] G. Bao, H. Zhang, C. Wang, L. Yang, and Y. Zhang, "How likely do llms with cot mimic human reasoning?" in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 7831–7850.
- [13] D. Paul, R. West, A. Bosselut, and B. Faltings, "Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning," arXiv preprint arXiv:2402.13950, 2024.
- [14] J. Pfau, W. Merrill, and S. R. Bowman, "Let's think dot by dot: Hidden computation in transformer language models," arXiv preprint arXiv:2404.15758, 2024.
- [15] F. Barez, T.-Y. Wu, I. Arcuschin, M. Lan, V. Wang, N. Siegel, N. Collignon, C. Neo, I. Lee, A. Paren *et al.*, "Chain-of-thought is not explainability."
- [16] J. Pearl, Causality. Cambridge university press, 2009.
- [17] Y. Zhou and Z. Zhu, "Towards robust text classification: Mitigating spurious correlations with causal learning," arXiv e-prints, pp. arXiv– 2411, 2024.
- [18] S. Harsha Tanneru, D. Ley, C. Agarwal, and H. Lakkaraju, "On the hardness of faithful chain-of-thought reasoning in large language models," arXiv e-prints, pp. arXiv-2406, 2024.

- [19] E. Yee, A. Li, C. Tang, Y. H. Jung, R. Paturi, and L. Bergen, "Dissociation of faithful and unfaithful reasoning in llms," arXiv preprint arXiv:2405.15092, 2024.
- [20] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray et al., "Training language models to follow instructions with human feedback," Advances in neural information processing systems, vol. 35, pp. 27730–27744, 2022.
- [21] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," 2022.
- [22] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models," Feb. 2024. [Online]. Available: https://arxiv.org/abs/2402.03300v3
- [23] C. Zheng, S. Liu, M. Li, X.-H. Chen, B. Yu, C. Gao, K. Dang, Y. Liu, R. Men, A. Yang, J. Zhou, and J. Lin, "Group Sequence Policy Optimization," Jul. 2025. [Online]. Available: http://arxiv.org/abs/2507. 18071
- [24] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, X. Liu, H. Lin, Z. Lin, B. Ma, G. Sheng, Y. Tong, C. Zhang, M. Zhang, W. Zhang, H. Zhu, J. Zhu, J. Chen, J. Chen, C. Wang, H. Yu, Y. Song, X. Wei, H. Zhou, J. Liu, W.-Y. Ma, Y.-Q. Zhang, L. Yan, M. Qiao, Y. Wu, and M. Wang, "DAPO: An Open-Source LLM Reinforcement Learning System at Scale," May 2025. [Online]. Available: http://arxiv.org/abs/2503.14476
- [25] Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin, "Understanding R1-Zero-Like Training: A Critical Perspective," Mar. 2025. [Online]. Available: http://arxiv.org/abs/2503.20783
- [26] X. Chu, H. Huang, X. Zhang, F. Wei, and Y. Wang, "GPG: A Simple and Strong Reinforcement Learning Baseline for Model Reasoning," May 2025. [Online]. Available: http://arxiv.org/abs/2504.02546
- [27] Z. Chu, J. Chen, Q. Chen, W. Yu, T. He, H. Wang, W. Peng, M. Liu, B. Qin, and T. Liu, "A survey of chain of thought reasoning: Advances, frontiers and future," arXiv preprint arXiv:2309.15402, 2023.
- [28] F. Yu, H. Zhang, P. Tiwari, and B. Wang, "Natural language reasoning, a survey," ACM Comput. Surv., may 2024, just Accepted. [Online]. Available: https://doi.org/10.1145/3664194
- [29] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," arXiv preprint arXiv:2203.11171, 2022.
- [30] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," arXiv preprint arXiv:2305.10601, 2023.
- [31] Y. Yao, Z. Li, and H. Zhao, "Beyond chain-of-thought, effective graph-of-thought reasoning in large language models," arXiv preprint arXiv:2305.16582, 2023.
- [32] Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, and C. Callison-Burch, "Faithful chain-of-thought reasoning," arXiv preprint arXiv:2301.13379, 2023.
- [33] R. Vacareanu, A. Pratik, E. Spiliopoulou, Z. Qi, G. Paolini, N. A. John, J. Ma, Y. Benajiba, and M. Ballesteros, "General purpose verification for chain of thought prompting," arXiv preprint arXiv:2405.00204, 2024.
- [34] Anthropic, "Introducing Claude 4," https://www.anthropic.com/news/ claude-4, 2025.
- [35] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, L. Marris, S. Petulla, C. Gaffney et al., "Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities," Jul. 2025. [Online]. Available: http://arxiv.org/abs/2507.06261
- [36] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu, "Qwen3 technical report," arXiv preprint arXiv:2505.09388, 2025.
- [37] xAI, "Grok 4," https://x.ai/news/grok-4, 2025.
- [38] OpenAI, "Introducing OpenAI o3 and o4-mini," https://openai.com/index/introducing-o3-and-o4-mini/, 2025.
- [39] —, "Introducing GPT-5," https://openai.com/index/introducing-gpt-5/, 2025.
- [40] Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen, "MMLU-Pro: A More Robust and Challenging

- Multi-Task Language Understanding Benchmark," Nov. 2024. [Online]. Available: http://arxiv.org/abs/2406.01574
- [41] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," Nov. 2023. [Online]. Available: http://arxiv.org/abs/2311.12022
- [42] B. Patel, S. Chakraborty, W. A. Suttle, M. Wang, A. S. Bedi, and D. Manocha, "AIME: AI System Optimization via Multiple LLM Evaluators," Oct. 2024. [Online]. Available: http://arxiv.org/abs/2410. 03131
- [43] Y. Qin, X. Li, H. Zou, Y. Liu, S. Xia, Z. Huang, Y. Ye, W. Yuan, H. Liu, Y. Li, and P. Liu, "O1 replication journey: A strategic progress report part 1," 2024. [Online]. Available: https://arxiv.org/abs/2410.18982
- [44] Y. Zhang, S. Wu, Y. Yang, J. Shu, J. Xiao, C. Kong, and J. Sang, "o1-coder: an o1 replication for coding," 2024. [Online]. Available: https://arxiv.org/abs/2412.00154
- [45] J. Wang, M. Fang, Z. Wan, M. Wen, J. Zhu, A. Liu, Z. Gong, Y. Song, L. Chen, L. M. Ni, L. Yang, Y. Wen, and W. Zhang, "Openr: An open source framework for advanced reasoning with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2410.09671
- [46] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto, "S1: Simple test-time scaling," Mar. 2025. [Online]. Available: http://arxiv.org/abs/2501.19393
- [47] P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, and M. Farajtabar, "The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity," arXiv preprint arXiv:2506.06941, 2025.
- [48] Y. Yue, Z. Chen, R. Lu, A. Zhao, Z. Wang, Y. Yue, S. Song, and G. Huang, "Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?" 2025. [Online]. Available: https://arxiv.org/abs/2504.13837
- [49] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," in *Proceedings of the AAAI conference* on artificial intelligence, vol. 33, 2019, pp. 3027–3035.
- [50] Z. Jin, Y. Chen, F. Leeb, L. Gresele, O. Kamal, L. Zhiheng, K. Blin, F. G. Adauto, M. Kleiman-Weiner, M. Sachan et al., "Cladder: Assessing causal reasoning in language models," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [51] Z. Jin, J. Liu, Z. Lyu, S. Poff, M. Sachan, R. Mihalcea, M. Diab, and B. Schölkopf, "Can large language models infer causation from correlation?" arXiv preprint arXiv:2306.05836, 2023.
- [52] E. Kıcıman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," arXiv preprint arXiv:2305.00050, 2023.
- [53] M. Zečević, M. Willig, D. S. Dhami, and K. Kersting, "Causal parrots: Large language models may talk causality but are not causal," *Transactions on Machine Learning Research*, 2023.
- [54] Q. Chen, L. Qin, J. Liu, D. Peng, J. Guan, P. Wang, M. Hu, Y. Zhou, T. Gao, and W. Che, "Towards reasoning era: A survey of long chain-of-thought for reasoning large language models," arXiv preprint arXiv:2503.09567, 2025.
- [55] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.
- [56] J. Angrist and G. Imbens, "Identification and estimation of local average treatment effects," 1995.
- [57] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.
- [58] OpenAI, "ChatGPT," https://chat.openai.com/, 2022.
- [59] —, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.
- [60] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.
- [61] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," arXiv preprint arXiv:2412.15115, 2024.
- [62] Q. Team, "Qwq-32b: Embracing the power of reinforcement learning," March 2025. [Online]. Available: https://qwenlm.github. io/blog/qwq-32b/

- [63] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," 2023.
- [64] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano et al., "Training verifiers to solve math word problems," arXiv preprint arXiv:2110.14168, 2021.
- [65] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," arXiv preprint arXiv:2305.20050, 2023.
- [66] O. Tafjord, B. D. Mishra, and P. Clark, "Proofwriter: Generating implications, proofs, and abductive statements over natural language," arXiv preprint arXiv:2012.13048, 2020.
- [67] S. Han, H. Schoelkopf, Y. Zhao, Z. Qi, M. Riddell, L. Benson, L. Sun, E. Zubova, Y. Qiao, M. Burtell et al., "Folio: Natural language reasoning with first-order logic," arXiv preprint arXiv:2209.00840, 2022.
- [68] H. Liu, J. Liu, L. Cui, Z. Teng, N. Duan, M. Zhou, and Y. Zhang, "Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.
- [69] J. Wu, L. Yang, Z. Wang, M. Okumura, and Y. Zhang, "Cofca: A step-wise counterfactual multi-hop qa benchmark," arXiv preprint arXiv:2402.11924, 2024.
- [70] J. Schulman, "Reinforcement learning from human feedback: Progress and challenges," in *Berkley Electrical Engineering and Computer Sciences. URL: https://eecs. berkeley. edu/research/colloquium/230419 [accessed 2023-11-15]*, 2023.
- [71] Y. Yang, E. Chern, X. Qiu, G. Neubig, and P. Liu, "Alignment for honesty," arXiv preprint arXiv:2312.07000, 2023.
- [72] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.
- [73] M. Luo, S. Tan, J. Wong, X. Shi, W. Y. Tang, M. Roongta, C. Cai, J. Luo, L. E. Li, R. A. Popa, and I. Stoica, "Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl," https://pretty-radio-b75.notion.site/ DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2, 2025, notion Blog.
- [74] T. Chu, Y. Zhai, J. Yang, S. Tong, S. Xie, D. Schuurmans, Q. V. Le, S. Levine, and Y. Ma, "Sft memorizes, rl generalizes: A comparative study of foundation model post-training," arXiv preprint arXiv:2501.17161, 2025.
- [75] A. Wu, K. Kuang, M. Zhu, Y. Wang, Y. Zheng, K. Han, B. Li, G. Chen, F. Wu, and K. Zhang, "Causality for large language models," arXiv preprint arXiv:2410.15319, 2024.
- [76] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2410.05229
- [77] J. Deng, J. Chen, Z. Chen, D. Cheng, F. Bai, B. Zhang, Y. Min, Y. Gao, W. X. Zhao, and J.-R. Wen, "From trial-and-error to improvement: A systematic analysis of llm exploration mechanisms in rlvr," arXiv preprint arXiv:2508.07534, 2025.
- [78] Q. Chen, L. Qin, J. Liu, D. Peng, J. Guan, P. Wang, M. Hu, Y. Zhou, T. Gao, and W. Che, "Towards reasoning era: A survey of long chain-of-thought for reasoning large language models," 2025. [Online]. Available: https://arxiv.org/abs/2503.09567