Mitigating Hallucinations in Large Language Models via Causal Reasoning

Yuangang Li[1,*], Yiqing Shen[2,*], Yi Nian[1], Jiechao Gao[3], Ziyi Wang[4], Chenxiao Yu[1], Shawn Li[1], Jie Wang[3], Xiyang Hu[5,†], Yue Zhao[1,†]

[1]University of Southern California
[2]Johns Hopkins University
[3]Stanford University
[4]University of Maryland, College Park
[5]Arizona State University
{yuangang,yinian,cyu96374,lli15753,yzhao010}@usc.edu, yshen92@jhu.edu,
{jiechao,jiewang}@stanford.edu, zoewang@umd.edu, xiyanghu@asu.edu

Abstract

Large language models (LLMs) exhibit logically inconsistent hallucinations that appear coherent yet violate reasoning principles, with recent research suggesting an inverse relationship between causal reasoning capabilities and such hallucinations. However, existing reasoning approaches in LLMs, such as Chain-of-Thought (CoT) and its graph-based variants, operate at the linguistic token level rather than modeling the underlying causal relationships between variables, lacking the ability to represent conditional independencies or satisfy causal identification assumptions. To bridge this gap, we introduce causal-DAG construction and reasoning (CDCR-SFT), a supervised fine-tuning framework that trains LLMs to explicitly construct variable-level directed acyclic graph (DAG) and then perform reasoning over it. Moreover, we present a dataset comprising 25,368 samples (CausalDR), where each sample includes an input question, explicit causal DAG, graph-based reasoning trace, and validated answer. Experiments on four LLMs across eight tasks show that CDCR-SFT improves the causal reasoning capability with the state-of-the-art 95.33% accuracy on CLADDER (surpassing human performance of 94.8% for the first time) and reduces the hallucination on HaluEval with 10% improvements. It demonstrates that explicit causal structure modeling in LLMs can effectively mitigate logical inconsistencies in LLM outputs. Code is available at https://github.com/MrLYG/CDCR-SFT.

1 Introduction

Large language models (LLMs) may generate logically inconsistent hallucinations during reasoning, where their outputs appear coherent but contain logical inconsistencies, leading to suboptimal performance (Banerjee, Agarwal, and Singla 2024; Huang et al. 2025; Cheng et al. 2025). Recent studies point out a correlation between causal reasoning capabilities and these logical inconsistency hallucinations (Bagheri et al. 2024; Wang 2024; Liu et al. 2025), namely LLMs with stronger causal reasoning abilities typically exhibit fewer logically inconsistent hallucinations. This observation motivates the central research question of this work:

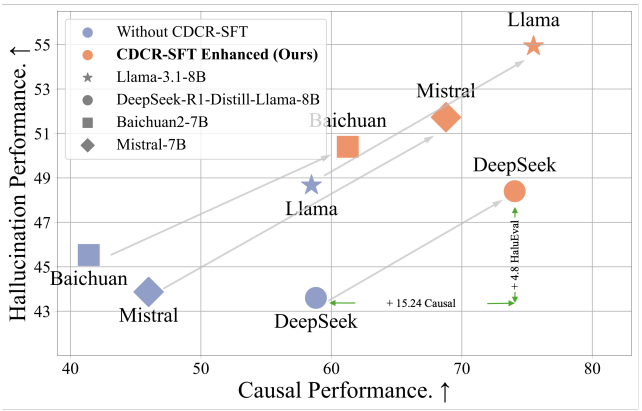


Figure 1: Average overall performance of our CDCR-SFT applied to four LLMs on the **causal reasoning** benchmarks (CLADDER and WIQA) and the **hallucination** benchmark (HaluEval). Orange symbols denote the LLMs enhanced by CDCR-SFT, demonstrating that CDCR-SFT significantly improves causal reasoning capabilities and reduces hallucinations.

"***Can we mitigate hallucinations by improving the causal reasoning capabilities of LLMs?***"

To answer this question, we must enhance LLMs' causal reasoning abilities. However, true causal reasoning requires LLMs to represent and manipulate a directed acyclic graph (DAG) that encodes conditional independence relationships, enables intervention queries, and satisfies causal identification assumptions (*i.e.*, exchangeability, consistency, positivity) (Hernan and Robins 2020) for identifying confounding effects. Existing structured-reasoning methods, including Chain-of-Thought (CoT) (Wei et al. 2022), Tree-of-Thought (ToT) (Yao et al. 2023), Graph-of-Thought (GoT) (Besta et al. 2024), and Diagram-of-Thought (DoT) (Zhang, Yuan, and Yao 2024), operate at the wrong level of abstraction, which models dependencies between linguistic tokens rather than causal relationships between variables (Bao et al. 2024; Fu et al. 2025; Luo, Zhang, and Li 2025). These meth-

*These authors contributed equally.
†Corresponding authors.

ods generate reasoning structures only at inference time through prompting, without any training signal to correct mis-specified causal relationships. Consequently, when an LLM incorrectly identifies A as causing B (when B actually causes A), or fails to recognize a confounding variable C that influences both, no gradient flows back to fix these fundamental errors (Wang et al. 2023; Yao et al. 2023; Besta et al. 2024). As a result, they cannot block spurious backdoor paths or guarantee counterfactual consistency, leaving LLMs still vulnerable to logically inconsistent hallucinations (Wang et al. 2023; Yao et al. 2023; Besta et al. 2024). The mathematical constraints further compound this problem. Causal relationships inherently form a DAG that encodes multiple interconnected variables with conditional dependencies and multiple pathways of influence. A linear chain or even a tree structure cannot adequately represent scenarios where a variable influences multiple outcomes simultaneously or where effects depend on the interaction of multiple causes, both fundamental characteristics of causal DAG. This structural mismatch means that prompt-only variants such as CoT, ToT, GoT, and DoT cannot, by design, supervise LLMs to learn causal edge semantics, limiting their ability to enforce conditional independencies required for true causal inference.

To address this gap, we propose **C**ausal-**D**AG **C**onstruction and **R**easoning (CDCR-SFT), a supervised fine-tuning framework that trains LLMs to first construct a variable-level causal DAG and then reason over that graph. The training of CDCR-SFT requires data with a causal DAG as well as the corresponding reasoning on top of that. Therefore, we introduce CausalDR (Causal-DAG and Reasoning), the first dataset specifically designed to train LLMs in simultaneous causal DAG construction and graph-based reasoning. Building upon the CLADDER dataset (Jin et al. 2023), which provides causal questions with a causal DAG, we develop an automated generation and validation pipeline using DeepSeek-R1 (DeepSeek-AI 2025). This pipeline ensures high-quality data generation through question-answer consistency checks. Each sample in CausalDR comprises (1) an input question, (2) a causal DAG that explicitly describes variables and their relationships, (3) a graph-based reasoning trace that navigates the causal structure, and (4) the final answer. As shown in Fig. 1, our experiments demonstrate that CDCR-SFT can address our research question by both improving causal reasoning capabilities and mitigating the logically inconsistent hallucinations across multiple benchmarks. This indicates that, rather than solely pursuing larger model sizes or more training data or longer cot, we can achieve more trustworthy LLMs by equipping them with structured reasoning capabilities that align with the underlying causal nature of real-world problems.

The major contributions of this work are three-fold. First, we introduce CDCR-SFT, a supervised fine-tuning framework that shifts how LLMs approach causal reasoning by moving from sequential CoT to DAG-based inference. It trains models to construct a causal DAG that properly encodes both causal directionality and conditional independence relationships, enabling them to perform structured reasoning over these graphs rather than being constrained by linear reasoning paths. Second, we present CausalDR, a dataset containing 25,368 high-quality samples for teaching LLMs to generate causal DAG construction and reason on top of the DAG. Third, we demonstrate that explicit causal structure modeling can not only improve causal reasoning but also mitigate hallucinations in LLMs.

2 Related Works

Reasoning and Causal Limitations in LLMs LLMs employ structured reasoning methods such as Chain-of-Thought (CoT) prompting, which generates intermediate steps alongside final answers (Wei et al. 2022); Self-Consistency (CoT-SC), which samples multiple reasoning chains for robustness; Tree-of-Thoughts (ToT), which branches into alternative solution paths (Yao et al. 2023); and Graph-of-Thoughts (GoT), which links subproblems as nodes in a simple graph (Besta et al. 2024). However, these methods treat inference as linear sequences or trees and cannot represent directed acyclic graph (DAG) needed for causal analysis, where edges denote cause–effect relations and support interventions and counterfactual reasoning. Benchmarks such as CausalBench show that LLMs struggle with intervention and counterfactual queries, failing to predict outcomes of hypothetical changes (Wang 2024), and synthetic tests confirm that models rely on surface text patterns rather than true cause–effect relations (Ma 2024).

Hallucination Reduction and Causal Supervised Fine-Tuning Complex reasoning tasks can exacerbate hallucinations in LLMs, as models often rely on surface-level correlations rather than true causal structure (Bagheri et al. 2024). Traditional mitigation—external knowledge checks or post-hoc filters—only corrects errors after generation and does not strengthen the model's internal inference process (Wang 2024). Recent studies have demonstrated that task-specific fine-tuning significantly improves LLM performance on specialized benchmarks (Han et al. 2024; Liu et al. 2025). In particular, supervised fine-tuning (SFT) with low-rank adapters (LoRA) (Hu et al. 2022) reshapes internal reasoning by training models on structured targets. In this study, we extend this paradigm by using the CausalDR dataset's annotated DAG and stepwise reasoning to teach the model to first construct a causal graph and then perform graph-based inference, thereby reducing hallucinations and improving consistency.

3 Methods

3.1 CDCR-SFT

CDCR-SFT is a supervised fine-tuning framework that trains LLMs to explicitly perform causal reasoning through causal-DAG construction and reasoning. Specifically, LLMs learn to construct a causal DAG by identifying causal variables from input queries, then perform structured reasoning over the DAG, and finally generate answers, as shown in Fig. 2.

Existing structured reasoning methods, such as CoT, ToT, and GoT, generally produce reasoning paths at the linguis-

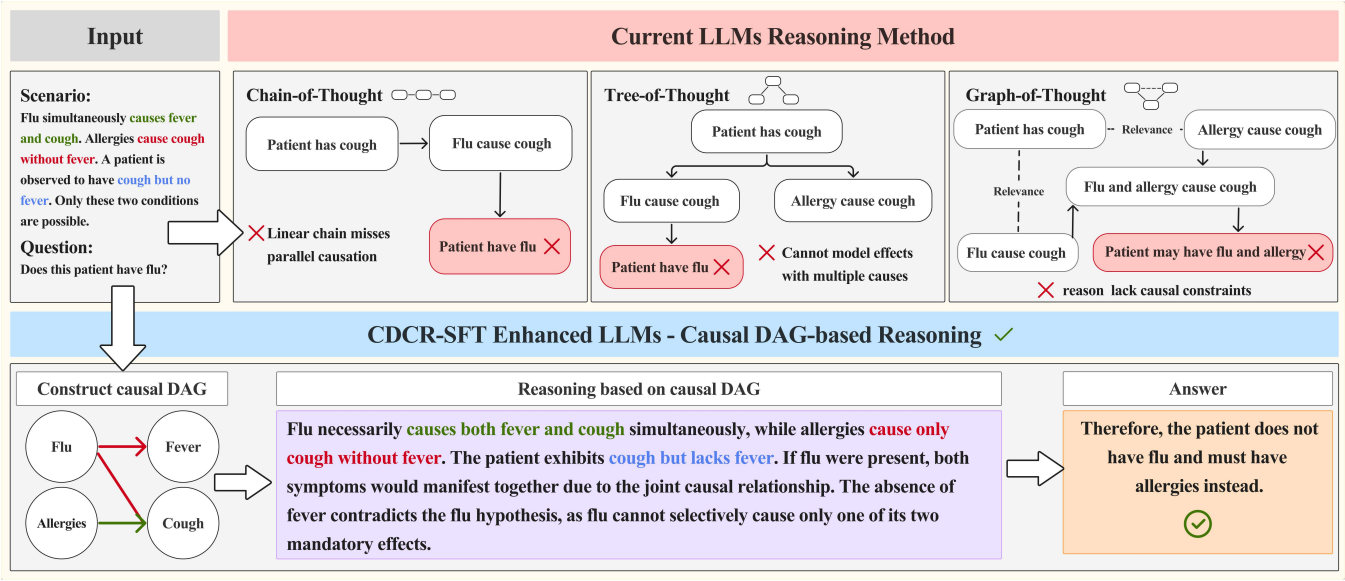


Figure 2: Comparison of reasoning approaches: Existing methods (CoT, ToT, GoT) operate at linguistic/semantic levels without explicit causal structure; Our CDCR-SFT constructs a variable-level causal DAG with directed edges representing causal relationships, enabling principled causal inference through graph-based reasoning.

tic token or semantic levels without modeling the underlying causal structures among variables. Table 1 provides a detailed comparison of key capabilities between our proposed CDCR-SFT framework and existing reasoning methods. Mathematically, CoT generates reasoning paths as linear reasoning sequences $S_{\text{CoT}} = (p_1, \ldots, p_n, y)$, ToT forms branching reasoning trees $S_{\text{ToT}} = \text{Tree}(p_1, \ldots, p_n, y)$, and GoT creates semantic-level reasoning graphs $S_{\text{GoT}} = \text{Graph}(p_1, \ldots, p_n, y)$. CDCR-SFT outputs a DAG-based

Aspect	CDCR-SFT (ours)	CoT	ToT	GoT
Reasoning aligned with causal relationships	✓	×	×	×
Explicit causal training signal	✓	×	×	×
Supports multi-parent causes	✓	×	×	× (no causal)
Captures conditional independencies	✓	×	×	×
Captures interventions	✓	×	×	×
Captures counterfactuals	✓	×	×	×
Effective hallucination mitigation	✓	×	×	×

Table 1: Comparison of key capabilities between CDCR-SFT and existing reasoning methods.

reasoning process $S_{CDCR\text{-}SFT} = (G, P, y)$, where $G = (V, E)$ denotes the causal DAG encoding causal directionality and conditional independence relationships, $P = (p_1(G), \ldots, p_n(G))$ represents reasoning steps that adhere strictly to causal structures in G, and y is the final inferred answer. Specifically, in the textual encoding of the causal DAG G, each causal variable is clearly represented as a node described in natural language, including detailed descriptions of the primary events. The causal relationships among these variables are encoded as directed edges, explicitly indicating directional influences. An illustrative example of textual DAG encoding is provided in Fig. 3.

Causal DAG

Nodes:
Node ID: X; Node Name: Flu ; Node Description: Virus causing fever and cough simultaneously.
Node ID: V2; Node Name: Allergy; Node Description: Immune reaction causing cough only, without fever.
Node ID: V3; Node Name: Fever; Node Description: Elevated body temperature, indicative of infection.
Node ID: Y; Node Name: Cough; Node Description: Reflex clearing the airway, triggered by various causes.
Edges:
Node: X; Inputs: N/A; Outputs: ['V3', 'Y'] Node: V2; Inputs: N/A; Outputs: ['Y']
Node: V3; Inputs: ['X']; Outputs: N/A Node: Y; Inputs: ['X', 'V2']; Outputs: N/A

Figure 3: Textual representation of the causal DAG in Fig. 2.

3.2 Dataset Construction

Causal-DAG and Reasoning (CausalDR) Dataset To train LLMs in simultaneous causal DAG construction and graph-based reasoning, we require datasets explicitly providing supervision for both. However, existing causal datasets (Gordon, Kozareva, and Roemmele 2012; Tandon et al. 2019; Du et al. 2022) either omit explicit causal relationships altogether or, as exemplified by CLADDER (Jin et al. 2023), offer mathematically rigorous yet semantically sparse causal graphs and algebraic formulations, lacking clear natural-language reasoning paths linking structures to answers (A CLADDER example is provided in Appx. A.1).

We introduce **CausalDR**, the first large-scale annotated dataset explicitly designed for supervised fine-tuning of LLMs in simultaneous causal DAG construction and structured causal reasoning. Each training sample in CausalDR consists of: (1) an instruction specifying the task, (2) an input question or scenario, and (3) a coherent output comprising three components: a text-based causal DAG G, a reasoning path $P == (p_1(G), \ldots, p_n(G))$ based on G, and a final answer y derived through structured inference (A detailed example see Appx. A.2).

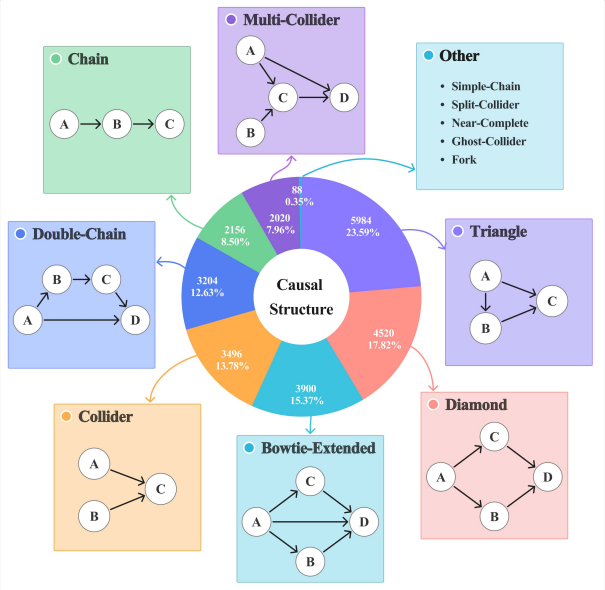


Figure 4: Proportional Distribution of 12 Canonical Causal DAG Structures in the CausalDR Dataset.

We construct CausalDR based on the CLADDER dataset (Jin et al. 2023), partitioning it into training and test sets based on unique identifiers (`graph_id` and `story_id`) to prevent information leakage. And then using the DeepSeek-R1 (DeepSeek-AI 2025) (temperature=0.6, max tokens=8192, details in Appx. A.3), we developed an automated pipeline (pseudocode provided in Appx. A.5) to generate and validate high-quality training samples for the CausalDR dataset. Specifically, we designed a prompt (details in Appx. A.4) that contains a mathematically accurate causal DAG expressed in formal notation, instructing DeepSeek-R1 to produce JSON-formatted outputs for each CLADDER sample. Each output explicitly: (1) causal nodes with clear semantic descriptions, and causal edges specifying incoming and outgoing relationships, (2) a step-by-step reasoning path that explicitly references the constructed causal DAG, and (3) the final inferred answer. To ensure quality, we implemented a validation mechanism comparing model-generated answers against the original ground-truth answers provided by CLADDER. If a generated answer did not match the ground-truth after multiple validation attempts, the sample was manually reviewed or discarded. Through this process, we obtained a high-quality dataset of 6,357 validated samples.

To further enhance dataset diversity and generalization, we introduced a Causal DAG Augmentation technique. Specifically, given an original causal DAG $G = (V, E)$, we randomly permuted the order of causal nodes and edges using permutation functions $\pi_v(\cdot)$ and $\pi_e(\cdot)$, respectively, to create diverse augmented variants: $V_{\text{aug}} = \pi_v(V), \quad E_{\text{aug}} = \pi_e(E), \quad G_{\text{aug}} = (V_{\text{aug}}, E_{\text{aug}})$. We applied this permutation procedure four times per original DAG G, each time pairing the permuted DAG G_{aug} with the original reasoning path P and answer y. This expanded the initial dataset from 6,357 samples to 25,368 augmented training examples.

Fig. 4 shows the proportional distribution of the 12 canonical causal DAG structures within the CausalDR dataset. These structures cover diverse causal configurations, including simple Chains (e.g., Chain, Double-Chain), Confounding structures (e.g., Triangle, Fork), Colliders (e.g., Collider, Multi-Collider), and more intricate multi-path interactions (e.g., Diamond, Bowtie-Extended). The diverse representation of these key causal mechanisms enables effective generalization of causal reasoning capabilities in large language models.

Auxiliary Instruction following Data To prevent the model from over-focusing on the causal task and degrading the linguistic generalization ability, we randomly select 10,000 Alpaca (Taori et al. 2023) examples and mix them with the CausalDR dataset during supervised fine-tuning to ensure the overall linguistic ability and generalization performance of the model.

3.3 Supervised Fine-tuning Procedure

During supervised fine-tuning, LLM learns to generate the structured causal DAG inference sequence $S_{CDCR\text{-}SFT} = (G, P, y)$. The optimization objective is formulated as a negative log-likelihood loss: $\mathcal{L}_{\text{CDCR-SFT}} = -\sum_{t=1}^{|S|} \log P(s_t \mid s_{<t}, X)$, where s_t denotes the t-th token in the ground-truth sequence S, and $s_{<t}$ represents all tokens before position t.

Critically, whenever the model-generated sequences deviate from the ground-truth causal DAG structure—such as introducing reversed causal edges, omitting essential causal variables, or adding extraneous causal relationships—explicit gradient signals immediately correct these inaccuracies. This supervision ensures that the model internalizes correct causal directionality, conditional independence properties, and intervention semantics required for accurate causal reasoning.

For computational efficiency, we applied Low-Rank Adaptation (LoRA) (Hu et al. 2022) during fine-tuning, updating only a small number of low-rank parameters inserted into each layer, while freezing the original pretrained LLM parameters. Through this fine-tuning procedure, CDCR-SFT trains the model to construct an accurate causal DAG and perform structured reasoning explicitly constrained by the causal relationships defined in these graphs, thereby improving the logical consistency of the LLM outputs and mitigating hallucinations.

4 Experiments

4.1 Experimental Setup

Base LLMs and Reasoning Methods We select four pretrained LLMs for evaluation: (1) Llama-3.1-8B-Instruct (Grattafiori et al. 2024), (2) DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI 2025), (3) Baichuan2-7B-Chat (Baichuan 2023), and (4) Mistral-7B-Instruct-v0.2 (Jiang et al. 2023). Additionally, we compare our CDCR-SFT method against five baseline reasoning approaches: Zero-shot-CoT(**CoT**)(Kojima et al. 2023), Chain-of-Thought Self-Consistency (**CoT-SC**)(Wang et al.

2023), Causal Chain-of-Thought (**CausalCoT**)(Jin et al. 2023), Tree-of-Thoughts (**ToT**)(Yao et al. 2023), and Graph-of-Thoughts (**GoT**) (Besta et al. 2024). Detailed descriptions of each baseline are provided in Appx. B.1.

Datasets We conduct experiments on 3 distinct datasets (Cladder (Jin et al. 2023), WIQA (Tandon et al. 2019), and HaluEval (Li et al. 2023)) to evaluate models' causal reasoning and hallucinations performance.

Dataset	#	Subtasks	Evaluation Focus
CLADDER	1,906	Rung 1, Rung 2, Rung 3	Causal reasoning; Causal DAG quality
WIQA	212	INPARA, EXOGENOUS	Causal reasoning
HaluEval	1,500	Dialogue, QA, Summarization	Hallucination

Table 2: Summary of datasets used in experiments.

CLADDER (Jin et al. 2023): A benchmark dataset evaluating LLMs' causal reasoning at three levels: Rung 1 (Association, observational correlations), Rung 2 (Intervention, active manipulation effects), and Rung 3 (Counterfactual, hypothetical "what-if" scenarios). Following preprocessing (see section 3.2), CLADDER is split into training and test sets by `graph_id` and `story_id` to avoid data leakage. To further ensure test data quality, we perform an additional validation step (details in Appx. B.2).

WIQA (Tandon et al. 2019): A challenging dataset for evaluating LLMs' causal reasoning capabilities. We focus on two perturbation types: in-paragraph (INPARA)—changes within the text that test causal chain reconstruction, and out-of-paragraph (EXOGENOUS)—external changes assessing the model's reasoning about external influences. We excluded irrelevant (no-effect) perturbations, as these modifications are unrelated to the original causal chain and do not effectively reflect a model's true causal reasoning capability. For efficiency and representativeness, we sampled 106 questions per subtask (total 212; 95% confidence, ±7% margin). Questions were systematically reformulated to reduce ambiguity (see Appx. B.3).

HaluEval (Li et al. 2023): A benchmark for evaluating models' hallucination across three NLP tasks: (1) Knowledge-grounded Dialogue (Dialogue), (2) Question Answering (QA), and (3) Text Summarization (Summarization). Each task includes paired examples, consisting of hallucinated samples (incorrect or unverifiable information) and corresponding factual samples. For our experiments, we randomly sample 500 pairs per task (total 1,500 pairs).

Evaluation Metrics We adopt two primary metrics to clearly evaluate the models' causal reasoning and hallucination reduction: (1) **Accuracy**: measures correctness in causal reasoning (CLADDER, WIQA) and hallucination (HaluEval) tasks. (2) **Causal DAG Quality**: evaluates *Node Score* (correct causal nodes), *Edge Score* (correct causal edges), and *Structural Score* (overall graph correctness, including directionality and completeness). Causal DAG is scored using GPT-4o-mini (Hurst et al. 2024), with detailed scoring criteria and evaluation procedures provided in Appx. B.4.

Implementation Details We perform LoRA fine-tuning on A40x4 GPUs using the LLaMA-Factory library (Zheng et al. 2024) with default hyperparameters. Fine-tuned models use vLLM (Kwon et al. 2023) on the same GPUs for inference. Base model inference is conducted through external platforms: DeepInfra for Llama-3.1-8B and Mistral-7B, and Baidu-Qianfan/OpenRouter for Baichuan2-7B and DeepSeek-R1-Distill-Llama-8B, with 200 concurrent threads. The inference temperature is set to 0.0 except for DeepSeek (0.6, following (DeepSeek-AI 2025)), CoT-SC (0.7, following (Wang et al. 2023)), and GoT (1.0, following (Besta et al. 2024)). Our method and CoT-based approaches utilize a unified three-step instruction, while CausalCoT, ToT, and GoT follow their original structured prompting(Jin et al. 2023; Yao et al. 2023; Besta et al. 2024). All reported results are averaged over 3 experimental runs. Complete implementation details, prompts, and configurations are provided in our available code.

4.2 Main Results and Analysis

Causal Reasoning Performance Table 3 reports the causal reasoning performance of our proposed CDCR-SFT method compared with five baseline methods (CoT, CoT-SC, CausalCoT, ToT, and GoT) across four different LLMs on two representative causal reasoning benchmarks: CLADDER and WIQA.

On the CLADDER benchmark, our CDCR-SFT consistently achieves improvements across all three causal reasoning levels (Rung 1: Association, Rung 2: Intervention, and Rung 3: Counterfactual). Specifically, with the Llama-3.1-8B-Instruct model, our method reaches an overall accuracy of 95.33%, surpassing the strongest baseline (CoT-SC: 72.88%) by an absolute margin of 22.45 percentage points. Remarkably, at the most challenging Counterfactual reasoning level (Rung 3), CDCR-SFT achieves a particularly large improvement of 27.75 percentage points, improving accuracy from 65.31% (CoT-SC) to 93.06%. More importantly, our approach is the first to surpass the human-level benchmark performance (94.8%) (Yu et al. 2025) on CLADDER. Similar consistent performance gains are also observed for the DeepSeek-R1-Distill-Llama-8B (74.29% to 92.44%), Baichuan2-7B-Chat (52.26% to 72.51%), and Mistral-7B-Instruct-v0.2 (59.60% to 92.76%) models.

On the WIQA benchmark, CDCR-SFT again achieves consistent improvements over all baseline methods. Taking the Llama-3.1-8B-Instruct model as an example, the overall accuracy is improved from the best baseline (CoT-SC: 52.36%) to 55.66%. Similar improvements are consistently observed for the DeepSeek-R1-Distill-Llama-8B (52.83% to 55.66%), Baichuan2-7B-Chat (33.49% to 50.00%), and Mistral-7B-Instruct-v0.2 (41.51 to 44.81%) models.

These consistent gains across multiple causal reasoning tasks (CLADDER and WIQA) and diverse model architectures—from instruction-tuned models (Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.2) to distilled variants (DeepSeek-R1-Distill-Llama-8B) and smaller-scale models (Baichuan2-7B-Chat)—reflect that the benefits of CDCR-SFT originate primarily from its explicit modeling of causal structures reasoning. Unlike conventional meth-

Method	Cladder (%)↑				WIQA (%)↑			HaluEval (%)↑			
	Rung1	Rung2	Rung3	*overall.*	INPARA	EXOGENOUS	*overall.*	Dialogue	QA	Summarization	*overall.*
Llama-3.1-8B											
CausalCoT	70.90	72.82	57.46	65.90	48.11	33.96	41.04	56.40	42.60	56.20	51.73
CoT	69.07	82.06	57.33	66.95	54.72	45.28	50.00	50.60	39.80	55.60	48.67
CoT-SC	72.87	88.13	65.31	72.88	60.38	44.34	52.36	43.60	34.00	52.60	43.40
ToT	71.17	79.16	64.79	70.20	56.60	45.28	50.94	52.80	42.00	58.00	50.93
GoT	61.21	76.78	58.90	63.38	55.66	47.17	51.42	50.20	43.40	50.20	47.93
CDCR-SFT (Ours)	**98.30**	**93.93**	**93.06**	**95.33**	**64.20**	**47.20**	**55.66**	**60.80**	**44.80**	**59.20**	**54.93**
DeepSeek-R1-Distill-Llama-8B											
CausalCoT	74.97	68.87	59.03	67.37	52.83	50.94	51.89	47.60	40.00	51.80	46.47
CoT	73.92	76.78	53.27	66.21	55.66	47.17	51.42	42.00	40.80	48.00	43.60
CoT-SC	77.98	88.13	63.74	74.29	51.89	43.40	47.64	33.60	41.40	32.20	35.73
ToT	70.34	80.62	66.18	71.14	56.60	44.34	50.47	39.40	43.80	40.20	41.13
GoT	75.23	80.97	57.63	69.43	55.66	50.00	52.83	**53.40**	41.00	50.60	48.33
CDCR-SFT (Ours)	**94.89**	**90.50**	**90.97**	**92.44**	**56.60**	**54.72**	**55.66**	48.60	**44.40**	**52.60**	**48.53**
Baichuan2-7B											
CausalCoT	50.46	46.44	51.70	50.16	22.64	27.36	25.00	45.80	48.40	43.20	45.80
CoT	49.67	62.01	48.56	51.68	34.91	27.36	31.13	44.20	46.60	45.80	45.53
CoT-SC	51.38	61.21	48.69	52.26	36.79	30.19	33.49	47.80	45.80	47.40	47.00
ToT	49.67	58.05	50.65	51.73	34.91	20.75	27.83	44.80	45.80	48.01	46.20
GoT	51.11	58.84	49.61	52.05	31.13	30.19	30.66	41.80	43.80	40.80	42.13
CDCR-SFT (Ours)	**71.04**	**75.20**	**72.64**	**72.51**	**50.00**	**50.00**	**50.00**	**50.60**	**49.60**	**51.00**	**50.40**
Mistral-7B											
CausalCoT	51.11	63.06	45.16	51.10	38.68	27.36	33.02	45.20	47.80	41.60	44.87
CoT	52.29	59.63	53.53	54.25	40.57	34.91	37.74	43.60	44.20	43.80	43.87
CoT-SC	56.75	66.75	58.90	59.60	42.45	38.68	40.57	44.40	45.20	44.00	44.53
ToT	50.46	56.20	50.39	51.57	42.45	32.08	37.26	47.00	42.80	46.60	45.47
GoT	50.85	63.85	56.15	55.56	42.45	40.57	41.51	47.60	46.20	46.80	46.87
CDCR-SFT (Ours)	**94.23**	**94.46**	**90.45**	**92.76**	**43.40**	**46.23**	**44.81**	**53.40**	**48.20**	**53.60**	**51.73**

Table 3: Performance comparison between our proposed CDCR-SFT and baseline reasoning methods on causal reasoning benchmarks (CLADDER and WIQA) and hallucination benchmark (HaluEval) across four different LLMs. Accuracy (%) is reported for overall benchmarks and subtasks; best results per model and task highlighted in bold.

ods that perform token-level or semantic-level reasoning, our approach trains LLMs to explicitly construct and reason over causal DAG, thus embedding a stronger inductive bias aligned with causal inference principles. Consequently, the models internalize improved representations of conditional independencies, intervention semantics, and causal directionality, facilitating more robust generalization across causal reasoning scenarios and tasks of varying complexity.

Hallucination Reduction Table 3 further reports the hallucination reduction performance of our proposed CDCR-SFT method across four different LLMs, evaluated on the HaluEval benchmark comprising three typical tasks: Dialogue, QA, and Summarization.

Our CDCR-SFT method consistently outperforms baseline reasoning methods in terms of overall accuracy on the HaluEval benchmark, demonstrating clear reductions in logical inconsistencies and hallucinations. Specifically, using the Llama-3.1-8B model, CDCR-SFT achieves an overall accuracy of 54.93%, significantly higher than the strongest baseline (CausalCoT: 51.73%) and substantially surpassing CoT-SC (43.40%) by over 11 percentage points. Particularly noteworthy is the Dialogue subtask, where accuracy improves from 43.60% (CoT-SC) to 60.80%, highlighting the effectiveness of our approach in mitigating hallucinations in complex interactive reasoning tasks.

Similar trends are evident for other evaluated LLMs. For instance, the DeepSeek improves from the strongest baseline (CausalCoT: 46.47%) to 48.40%, Baichuan improves from 47.00% (CoT-SC) to 50.40%, and Mistral shows accuracy improvement from the best baseline (GoT: 46.87%) to 51.73%. Importantly, these significant hallucination reductions are achieved without hallucination-focused supervision, indicating that the reduction naturally arises from enhanced causal reasoning capabilities learned by the model.

These empirical findings directly support our core hypothesis: explicitly improving the causal reasoning capabilities of LLMs inherently mitigates logically inconsistent hallucinations. The substantial and consistent hallucination reductions observed across diverse tasks and model architectures demonstrate that our CDCR-SFT method provides an effective and generalizable solution for enhancing the reliability and consistency of LLMs.

4.3 Causal DAG Construction Quality

CDCR-SFT is to enable LLMs to reason accurately based on a variable-level causal DAG. The quality of the generated DAG thus directly reflects the extent to which the model has internalized correct causal relationships and structured causal reasoning capabilities, including accurately capturing causal directionality, conditional independencies, and satisfying causal identification assumptions. We compare the Causal DAG generated using pre-trained LLMs versus the

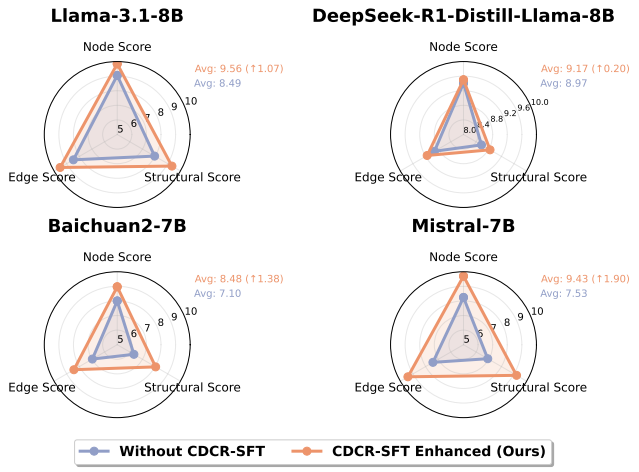


Figure 5: Comparison of causal DAG quality scores (Node, Edge, and Structural Scores) generated by pretrained LLMs versus those enhanced with CDCR-SFT, evaluated on the CLADDER dataset.

DAG produced by LLMs enhanced with our CDCR-SFT approach. Both employ the same prompt instructing models to generate Causal DAG. Fig. 5 indicates that CDCR-SFT raises scores in each dimension for all models. For Llama-3.1-8B, the overall average increases from 8.49 to 9.56, with the largest rise in Structural Score (7.96 to 9.33). DeepSeek-R1-Distill-Llama-8B shows a small increase from 8.97 to 9.17, chiefly in Edge Score (8.92 to 9.15). Baichuan2-7B advances from 7.10 to 8.48, with a 1.72-point gain in Structural Score. Mistral-7B displays the greatest progress, from 7.53 to 9.43, with gains over 2 points in both Edge Score and Structural Score. The significantly higher DAG quality achieved by CDCR-SFT over baseline methods validates that explicit DAG-based reasoning supervision enhances LLMs' capability to correctly represent and reason with causal structures, directly supporting improvements observed in causal reasoning tasks and hallucination reduction.

4.4 Ablation Study

We conduct an ablation study to assess whether the observed performance improvements originate specifically from our causal DAG construction and causal DAG-based reasoning strategy, or merely from the additional exposure to causal knowledge and examples provided during fine-tuning. Specifically, we compare three experimental conditions across all three benchmarks, reporting overall accuracy for CLADDER, WIQA, and HaluEval: (i) *Baseline*: the best-performing existing reasoning method per benchmark (selected from CoT, CoT-SC, ToT, GoT, and CausalCoT in Table 3); (ii) *CDCR-SFT-Ablated*: fine-tunes LLMs using only question-answer pairs from the CausalDR dataset, omitting causal DAG G construction and reasoning paths P, but retaining identical auxiliary instruction-following data; (iii) *CDCR-SFT*: our full proposed method, explicitly trained on causal DAG construction and DAG-based reasoning. All conditions maintain identical training configurations, including model architectures, hyperparameters, and data volumes, ensuring a fair comparison. Table 4 shows that

Method	Cladder (%)↑	WIQA (%)↑	HaluEval (%)↑
Llama-3.1-8B			
Baseline	72.88	52.36	51.73
CDCR-SFT-Ablated	87.25	49.06	44.97
CDCR-SFT (Ours)	**95.33**	**55.66**	**54.93**
DeepSeek-R1-Distill-Llama-8B			
Baseline	74.29	52.83	48.33
CDCR-SFT-Ablated	74.87	51.89	43.67
CDCR-SFT (Ours)	**92.44**	**55.66**	**48.53**
Baichuan2-7B			
Baseline	52.26	33.49	47.00
CDCR-SFT-Ablated	69.57	42.92	42.10
CDCR-SFT (Ours)	**72.51**	**50.00**	**50.40**
Mistral-7B			
Baseline	59.60	41.51	46.87
CDCR-SFT-Ablated	67.58	38.68	49.10
CDCR-SFT (Ours)	**92.76**	**44.81**	**51.73**

Table 4: Ablation study verifying the impact of explicit causal DAG-based reasoning, comparing baseline (best existing method), CDCR-SFT-Ablated (fine-tuned without causal DAG construction and reasoning), and our CDCR-SFT across three benchmarks on four LLMs.

fine-tuning models solely with causal question-answer pairs (CDCR-SFT-Ablated), without explicit causal DAG-based reasoning, consistently improves accuracy on the CLADDER benchmark (e.g., +14.4% on Llama-3.1-8B, +17.3% on Baichuan2-7B) but leads to performance degradation on the WIQA and HaluEval benchmarks compared to the Baseline. In contrast, our full method (CDCR-SFT), which learned causal DAG construction and causal DAG-based reasoning, consistently outperforms both the Baseline and CDCR-SFT-Ablated methods across all benchmarks and model architectures. These results confirm that the observed performance gains are attributable to structured causal reasoning rather than simply additional causal data exposure.

5 Conclusion

We propose CDCR-SFT, to shift how LLMs approach causal reasoning by moving from sequential CoT or graph variant to causal DAG-based reasoning. It trains models to construct a causal DAG that properly encodes both causal directionality and conditional independence relationships, enabling them to perform structured reasoning over the graph rather than being constrained by linear reasoning paths or causal-irrelevant graph reasoning. And we create the CausalDR dataset, containing 25,368 validated samples, provides high-quality supervision for LLMs to learn explicit causal DAG construction and graph-based reasoning. Our experiments across four LLMs on the CLADDER, WIQA, and HaluEval benchmarks demonstrate that CDCR-SFT significantly improves causal reasoning, achieving a state-of-the-art accuracy of 95.33% on CLADDER, surpassing human performance (94.8%) for the first time. Moreover, CDCR-SFT reduces hallucination in HaluEval by up to 11%, confirming that enhanced causal reasoning directly mitigates hallucinations. These results affirmatively

answer our research question: **improving the causal reasoning capabilities of LLMs can mitigate hallucinations**. In the future, rather than solely pursuing larger model sizes or more training data or longer cot, we can achieve more trustworthy LLMs by equipping them with structured reasoning capabilities that align with the underlying causal nature of real-world problems.

References

Bagheri, A.; Alinejad, M.; Bello, K.; and Akhondi-Asl, A. 2024. C^2P: Featuring Large Language Models with Causal Reasoning. arXiv:2407.18069.

Baichuan. 2023. Baichuan 2: Open Large-scale Language Models. *arXiv preprint arXiv:2309.10305*.

Banerjee, S.; Agarwal, A.; and Singla, S. 2024. Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746*.

Bao, G.; Zhang, H.; Wang, C.; Yang, L.; and Zhang, Y. 2024. How Likely Do LLMs with CoT Mimic Human Reasoning? *arXiv preprint arXiv:2402.16048*.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the CDCR conference on artificial intelligence*, volume 38, 17682–17690.

Cheng, F.; Li, H.; Liu, F.; van Rooij, R.; Zhang, K.; and Lin, Z. 2025. Empowering llms with logical reasoning: A comprehensive survey. *arXiv preprint arXiv:2502.15652*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Du, L.; Ding, X.; Xiong, K.; Liu, T.; and Qin, B. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv:2205.05849*.

Fu, J.; Ding, L.; Li, H.; Li, P.; Wei, Q.; and Chen, X. 2025. Unveiling and causalizing cot: A causal pespective. *arXiv preprint arXiv:2502.18239*.

Gordon, A.; Kozareva, Z.; and Roemmele, M. 2012. SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. 394–398. Montréal, Canada: Association for Computational Linguistics.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Han, Z.; Gao, C.; Liu, J.; Zhang, J.; and Zhang, S. Q. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Hernan, M.; and Robins, J. 2020. Causal inference: What if chapman hall/crc, boca raton.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Jin, Z.; Chen, Y.; Leeb, F.; Gresele, L.; Kamal, O.; Lyu, Z.; Blin, K.; Gonzalez, F.; Kleiman-Weiner, M.; Sachan, M.; and Schölkopf, B. 2023. CLadder: Assessing Causal Reasoning in Language Models. In *NeurIPS*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Li, J.; Cheng, X.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models.

Liu, X.; Xu, P.; Wu, J.; Yuan, J.; Yang, Y.; Zhou, Y.; Liu, F.; Guan, T.; Wang, H.; Yu, T.; et al. 2025. Large language models and causal inference in collaboration: A comprehensive survey. *Findings of the Association for Computational Linguistics: NAACL 2025*, 7668–7684.

Luo, H.; Zhang, J.; and Li, C. 2025. Causal graphs meet thoughts: Enhancing complex reasoning in graph-augmented llms. *arXiv preprint arXiv:2501.14892*.

Ma, J. 2024. Causal inference with large language model: A survey. *arXiv preprint arXiv:2409.09822*.

Tandon, N.; Mishra, B. D.; Sakaguchi, K.; Bosselut, A.; and Clark, P. 2019. Wiqa: A dataset for" what if..." reasoning over procedural text. *arXiv preprint arXiv:1909.04739*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171.

Wang, Z. 2024. CausalBench: A Comprehensive Benchmark for Evaluating Causal Reasoning Capabilities of Large Language Models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, 143–151. Bangkok, Thailand: Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Yu, L.; Chen, D.; Xiong, S.; Wu, Q.; Liu, Q.; Li, D.; Chen, Z.; Liu, X.; and Pan, L. 2025. CausalEval: Towards Better Causal Reasoning in Language Models. arXiv:2410.16676.

Zhang, Y.; Yuan, Y.; and Yao, A. C.-C. 2024. On the diagram of thought. *arXiv preprint arXiv:2409.10038*.

Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; and Ma, Y. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Bangkok, Thailand: Association for Computational Linguistics.

Supplementary Material

A Additional Methodological Details

A.1 Example of the Cladder Dataset

Our CausalDR dataset for supervised fine-tuning is based on the publicly available dataset, Cladder. The Cladder dataset was initially proposed to evaluate causal reasoning capabilities of large language models (LLMs). The dataset provides scenarios, questions, formal symbolic reasoning steps, and answers. Here, we showcase only the attributes directly relevant to our data construction process.

Below is a representative example from the Cladder dataset:

Original Cladder Dataset Sample

Scenario & Question: Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships: Husband has a direct effect on wife and alarm clock. Wife has a direct effect on alarm clock. For husbands that don't set the alarm and wives that don't set the alarm, the probability of ringing alarm is 8%. For husbands that don't set the alarm and wives that set the alarm, the probability of ringing alarm is 54%. For husbands that set the alarm and wives that don't set the alarm, the probability of ringing alarm is 41%. For husbands that set the alarm and wives that set the alarm, the probability of ringing alarm is 86%. For husbands that don't set the alarm, the probability of alarm set by wife is 74%. For husbands that set the alarm, the probability of alarm set by wife is 24%. If we disregard the mediation effect through wife, would husband positively affect alarm clock?

Reasoning:

Let X = husband; V2 = wife; Y = alarm clock.

X ->V2, X ->Y, V2 ->Y

$E[Y_{X=1,V2=0} - Y_{X=0,V2=0}]$

$\sum_{V2=v} P(V2 = v|X = 0) * [P(Y = 1|X = 1, V2 = v) - P(Y = 1|X = 0, V2 = v)]$

$P(Y = 1|X = 0, V2 = 0) = 0.08$

$P(Y = 1|X = 0, V2 = 1) = 0.54$

$P(Y = 1|X = 1, V2 = 0) = 0.41$

$P(Y = 1|X = 1, V2 = 1) = 0.86$

$P(V2 = 1|X = 0) = 0.74$

$P(V2 = 1|X = 1) = 0.24$

$0.74 * (0.86 - 0.41) + 0.24 * (0.54 - 0.08) = 0.32$

$0.32 > 0$

Answer: Yes

In this example, the scenario describes causal interactions among variables, the question focuses explicitly on causal effects, and the reasoning provides symbolic and probabilistic calculations. These symbolic representations serve as important references when identifying causal nodes and relationships for our structured data generation.

A.2 Example of the CausalDR Dataset(Ours)

CausalDR dataset used for supervised fine-tuning within the CDCR-SFT framework is presented in a clear, structured JSON format. Each sample contains three essential components: (1) a specific **instruction** explicitly guiding the model on what kind of task it should perform, (2) an **input** that provides the detailed scenario description and the specific causal reasoning question, and (3) an **output** demonstrating a clear, structured, and comprehensive reasoning process explicitly based on the provided causal graph.

Below, we provide a detailed example from CausalDR dataset to illustrate precisely how we formulate these training samples. Each sample not only explicitly represents the causal graph, inference path, and answer but also cohesively integrates these components into a natural and intuitive explanatory narrative. This structured presentation is specifically designed to encourage LLMs to internalize a systematic approach to causal reasoning, thereby significantly enhancing inference consistency and reducing hallucinations.

Fine-tuning Dataset Example (JSON)

```
{
   "instruction": "Given the question below, please construct a causal graph to analyze
the scenario. Then, based on the causal graph, provide a detailed explanation of the
step-by-step causal reasoning process. Finally, give the answer to the question based
on the causal graph and the reasoning process.",
   "input": "Imagine a self-contained, hypothetical world with only the following
```

conditions, and without any unmentioned factors or causal relationships: Husband has a direct effect on wife and alarm clock. Wife has a direct effect on alarm clock. For husbands that don't set the alarm and wives that don't set the alarm, the probability of ringing alarm is 8%. For husbands that don't set the alarm and wives that set the alarm, the probability of ringing alarm is 54%. For husbands that set the alarm and wives that don't set the alarm, the probability of ringing alarm is 41%. For husbands that set the alarm and wives that set the alarm, the probability of ringing alarm is 86%. For husbands that don't set the alarm, the probability of alarm set by wife is 74%. For husbands that set the alarm, the probability of alarm set by wife is 24%. If we disregard the mediation effect through wife, would husband positively affect alarm clock?",

"output": "Alright, let me first review your input. Next, I will build a causal graph from the information provided, defining each node and clarifying how they interact. After that, I will detail the steps of causal inference, describing how I move from the causal graph to the final answer. To ensure clarity, I will begin by presenting the causal graph's structure, the meaning of each node, and their connections. Then, I will illustrate the inference process, leading up to the result.

Causal Graph:

First, here is the section on the causal graph nodes. For each node, I will list its ID, Name, and provide a brief description.

Nodes:

Node ID: X
Node Name: Husband
Node Description: Binary variable representing whether the husband sets the alarm (1=sets, 0=doesn't set)
Node ID: Y
Node Name: Alarm Clock
Node Description: Binary outcome variable representing whether the alarm rings (1=rings, 0=doesn't ring)
Node ID: V2
Node Name: Wife
Node Description: Binary variable representing whether the wife sets the alarm (1=sets, 0=doesn't set)

Next, I will explain how these nodes are linked in the causal graph, showing how information flows between them. I will go through each node in turn, indicating which nodes feed into it and which nodes it influences.

Edges:

Node: X Inputs: N/A Outputs: ['V2', 'Y']
Node: Y Inputs: ['X', 'V2'] Outputs: N/A
Node: V2 Inputs: ['X'] Outputs: ['Y']

Based on the current provided input information, **the inference goal** is that "*Determine if Husband (X) has a positive direct effect on Alarm Clock (Y) when excluding mediation through Wife (V2)*"

Given this goal, the following describes the causal inference process.

Causal Inference Process: The causal graph shows Husband (X) directly influences both Wife (V2) and Alarm (Y), while Wife (V2) also directly affects Alarm (Y). To isolate X's direct effect on Y, we calculate the natural direct effect by holding V2 constant at its X=0 distribution. Using the provided probabilities: When X=0, P(V2=1)=0.74 and P(V2=0)=0.26. The direct effect compares Y probabilities under X=1 vs X=0 for these fixed V2 values. For V2=0: 0.41 (X=1) - 0.08 (X=0) = +0.33. For V2=1: 0.86 (X=1) - 0.54 (X=0) = +0.32. Weighting these differences by V2's X=0 distribution gives (0.26∗0.33) + (0.74∗0.32) = 0.32. This positive result confirms Husband has a direct positive effect on Alarm Clock when excluding the Wife-mediated pathway.

As a result of this causal inference process, I will reply with the answer yes."
}

A.3 Implementation Details of DeepSeek-R1

To construct the CausalDR dataset, we selected the reasoning large language model **DeepSeek-R1** (DeepSeek-AI 2025) due to its strong reasoning capabilities and low inference cost. The model was deployed and accessed via the **DeepInfra API**[1], an inference platform designed to streamline model integration. Following recommendations from DeepSeek-R1 (DeepSeek-AI 2025), we configured the temperature at **0.6** and set the maximum token length to **8192** during dataset generation.

[1]https://deepinfra.com

A.4 Prompt Template for Generating CausalDR Data

Below, we provide the detailed prompt template used to guide the DeepSeek-R1 model in generating detailed causal graphs and explicit natural language reasoning paths for each training sample.

Detailed Prompt Template for Structured Causal Inference Data Generation

You are an expert specializing in causal inference and graph theory. Your task is to analyze a reasoning problem, construct a structured causal graph, and generate a detailed causal inference process. Your output must be in JSON format.
You will receive:

- **Context & Question:** A single block (<context_question> ... </context_question>) that contains:
  - Scenario description
  - Constraints or rules
  - Any additional details
- **Reasoning:** This field contains the formal causal structure and mathematical reasoning needed to solve the problem. It includes:
  - Variable assignments (e.g., V1 = kraz, X = pexu)
  - Causal graph structure notation (e.g., V1→X, X→Y)
  - Probability calculations and mathematical steps required for the solution

Your task:

- Extract causal nodes and relationships from the provided input to construct a Causal Graph.
- Causal Graph is a Directed Acyclic Graph (DAG) that represents causal influences between different variables.
- Generate a structured causal reasoning process explaining how the conclusion is derived. In the `Causal Reasoning` field:
  - `goal`: A concise statement of the reasoning question.
  - `explanation`: A paragraph that:
    - Describes variables and causal edges
    - Explains causal influence propagation
    - Translates formal math into intuition
    - Justifies the final conclusion using probabilities

Return JSON Format:

```
{
  "Nodes": [
    {
      "id": "[DescriptiveVariableID]",
      "name": "[Variable Name]",
      "description": "[Detailed description of the causal variable]"
    }
    // ... more nodes if needed
  ],
  "Edges": [
    {
      "node": "[Same DescriptiveVariableID as in Nodes]",
      "inputs": ["List of all incoming causal nodes"],
      "outputs": ["List of all outgoing causal nodes"]
    }
    // ... Ensure that all Nodes are represented here.
  ],
  "Causal Reasoning": {
    "goal": "[Overarching question or objective]",
    "explanation": "[Step-by-step reasoning process]"
  },
  "Answer": "[yes/no]"
}
```

<context_question>[*Insert the given scenario description, constraints, and the specific question here.*</context_question>
<reasoning>[*Insert the symbolic causal graph structure, variable assignments, and mathematical reasoning steps provided by the Cladder dataset here.*]</reasoning>

A.5 Detailed Algorithm for Fine-tuning Dataset Construction

We provide a detailed pseudocode representation of our automated dataset construction pipeline in Algorithm 1. This algorithm clearly illustrates the structured process of generating high-quality fine-tuning data leveraging the DeepSeek-R1 model. Each step explicitly ensures data correctness, coherence, and suitability for supervised fine-tuning within our proposed framework.

Algorithm 1: CausalDR Dataset Construction

Input: Cladder training set $D_{\text{Cladder}} = \{(c_i, q_i, r_i, a_i)\}_{i=1}^{N}$, where:
c_i: scenario context;
q_i: causal inference question;
r_i: symbolic reasoning from Cladder dataset;
a_i: ground-truth answer.
LLM for causal reasoning: DeepSeek-R1; maximum attempts $K = 15$
Output: Fine-tuning dataset D_{CausalDR}
Initialize $D_{\text{CausalDR}} \leftarrow \emptyset$
for all $(c_i, q_i, r_i, a_i) \in D_{\text{Cladder}}$ **do**
    Construct structured prompt p_i from (c_i, q_i, r_i) (details in Appx. A.4)
    Set success $\leftarrow$ False, $k \leftarrow 0$
    while $\neg$success $\wedge\, k < K$ **do**
        $(G_i, P_i, y_i) \leftarrow$ DeepSeek-R1(p_i) ▷ structured JSON output
        if $y_i = a_i$ **then**
            Construct coherent inference paragraph S_i by explicitly integrating (G_i, P_i, y_i) into a natural explanatory narrative (See example output in Appx. A.2.).
            $D_{\text{CausalDR}} \leftarrow D_{\text{CausalDR}} \cup \{(c_i, q_i, S_i)\}$ ▷ S_i encapsulates G_i, P_i, y_i
            success $\leftarrow$ True
        end if
        $k \leftarrow k + 1$
    end while
end for
return D_{CausalDR}

Algorithm 1 systematically describes how we integrate DeepSeek-R1 to produce a coherent reasoning sequence (S) comprising structured causal graphs (G), explicit inference paths (P), and answers (y), thereby ensuring the resulting fine-tuning dataset addresses the challenge of hallucinations in LLM inference.

B Additional Experimental Details

B.1 Baselines Methods

To assess the effectiveness of our CDCR-SFT method, we compare it against 5 commonly used reasoning methods. While these methods have achieved some success in reasoning, they still have limitations in dealing with complex causal reasoning tasks, including the difficulty of effectively capturing causal relationships, as described in the Introduction section. Specifically, our comparative approach consists of:

- **Chain-of-Thought (CoT)** (Wei et al. 2022): CoT instructs the model to generate intermediate reasoning steps, helping models to solve complex problems by decomposing tasks into simpler sub-steps. We use zero-shot CoT without any few-shot prompting (Kojima et al. 2023), only the reasoning prompt, and refer to it as CoT in the following sections.
- **Chain-of-Thought Self-Consistency (CoT-SC)** (Wang et al. 2023): An improved version of CoT that samples multiple reasoning paths and selects the final answer based on consistency among these paths.
- **Causal Chain-of-Thought (CausalCoT)** (Jin et al. 2023): CausalCoT guides the model through defined steps, including causal graph extraction, formalization of queries, and calculation of counterfactual outcomes.
- **Tree-of-Thoughts (ToT)** (Yao et al. 2023): ToT organizes the reasoning process as a tree structure, enabling the model to explore multiple reasoning paths.
- **Graph-of-Thoughts (GoT)** (Besta et al. 2024): GoT organizes reasoning steps into a graph structure, modeling each reasoning step as a node and dependencies among these steps as edges, without explicitly modeling causal relationships.

B.2 Detailed Validation of Cladder Test Set

To ensure the quality and validity of our Cladder test set, we implemented a rigorous, two-step validation process involving both automated evaluation with DeepSeek-R1 and manual verification. Here, we describe this validation workflow in detail.

Step 1: Automated Validation via DeepSeek-R1. We constructed prompts that asked the DeepSeek-R1 model to verify the correctness of the provided reasoning and answer for each test sample. An example validation prompt is as follows:

Example Validation Prompt

```
You are an expert analyzing causal reasoning. Evaluate if the reasoning process and
answer are correct for this causal inference problem.
<context_question>...[Scenario and causal question here]...</context_question>
<reasoning>...[Provided symbolic reasoning steps]...</reasoning>
<proposed_answer>...[Provided answer]...</proposed_answer>
Provide a JSON response:
{
  "reasoning_valid": true/false,
  "reasoning_error": "Brief description of error if any, otherwise 'None'",
  "answer_correct": true/false,
  "correct_answer": "yes/no",
  "brief_explanation": "1-2 sentences explaining your assessment"
}
```

If either the reasoning or the answer was marked incorrect, these samples were flagged for further review.

Step 2: Manual Verification. Samples flagged as problematic by DeepSeek-R1 underwent manual review by domain experts to confirm the validity of the model's assessment. During this review, we carefully inspected reasoning accuracy and answer correctness, retaining only those samples unanimously confirmed as valid and logically sound.

Through this rigorous validation pipeline, we removed a total of 189 problematic samples, refining our test set down to 1,906 high-quality examples, suitable for rigorous causal reasoning evaluation.

The final validated Cladder test set, along with the validation scripts, will be publicly released to ensure reproducibility of our experiments.

B.3 WIQA Question Reformulation Procedure

Many original WIQA questions contained ambiguous phrasing or grammatical errors, potentially affecting evaluation results. An example of such ambiguity is:

Original Question (Ambiguous Example):
"Suppose less DNA available happens, how will it affect hurting the DNA to replicate properly?"
Options: A) more, B) less, C) no effect

To eliminate such issues, we reformulated all questions into a clear, standardized format, strictly matching the provided answer options. The improved version of the above question is:

Improved Question (Reformulated Example):
"Will having less available DNA cause more replication errors, fewer replication errors, or have no effect?"
Options: A) more, B) less, C) no effect

We used an automated approach employing the DeepSeek-R1 to rewrite each selected WIQA question. Below is the prompt we employed for the automatic rewriting:

Prompt Used for Question Reformulation

I have an English multiple-choice question with incorrect grammar and unclear meaning. I know that the correct answer is "{Answer choice}". Please help me rewrite this question so that:
- It is grammatically correct.
- It is logically clear and specific.
- It introduces no new information from outside the paragraph.
- It strictly preserves the original multiple-choice options:
A) more, B) less, C) no effect.
- The question must be rewritten exactly in the following format:
"Will [cause/change] cause more [effect], fewer [effect], or have no effect?"
This format must match the options exactly and avoid ambiguity. The rewritten question should clearly express the potential impact of a change on a specific outcome, and the options "more", "less", and "no effect" should directly correspond to the parts of the question.

Here is the background context:

Process steps:
{List of paragraph steps provided here}

Original question:
{Original problematic question provided here}

Options:
A) more
B) less
C) no effect

Correct answer: {Correct answer choice provided here}

Return your result strictly in the following JSON format:
{
"improved_question": "Your improved question here"
}

All 212 selected WIQA questions (106 from INPARA_EFFECT and 106 from EXOGENOUS_EFFECT) underwent this reformulation. A random subset of reformulated questions was manually reviewed to confirm grammatical correctness and logical clarity.

B.4 Details of LLM-based Evaluation for Causal Graph Quality

We employed an automatic evaluation approach utilizing an LLM (`gpt-4o-mini`) as a judge to assess the quality of the generated causal graphs. Specifically, given a causal reasoning context, a ground-truth causal graph, and a model-generated causal graph, the evaluator rated each generated graph along three dimensions: **Node Accuracy**, **Edge Accuracy**, and **Structural Fidelity**, assigning scores on a scale from 0 to 10. To ensure reproducibility, we set the inference temperature of `gpt-4o-mini` to 0.

Causal Graph Quality Scoring Criteria The detailed scoring criteria for these dimensions are presented in Table 5.

Prompt Template for LLM-based Evaluation The prompt template used for this automatic evaluation was as follows:

Prompt Template for LLM-based Evaluation

You are an expert evaluator specialized in assessing the quality of causal graph structures.

Your task:
Given a specific causal reasoning scenario (the problem context), along with a Ground Truth causal graph description (serving as the evaluation standard), your goal is to evaluate the quality of a **model-generated causal graph** (the evaluation target).

Table 5: Detailed scoring criteria for LLM-based causal graph quality evaluation.

Score	Node Accuracy	Edge Accuracy	Structural Fidelity
10	All nodes perfectly identified, no errors or omissions.	All edges (including directions) identified perfectly.	Structure perfectly matches the Ground Truth; fully reasonable.
9	All core nodes correctly identified; only minor discrepancies with non-critical nodes.	Nearly perfect; only one minor discrepancy on a non-critical edge.	Structure highly matches, minor irrelevant differences only.
8	Nearly all nodes correctly identified; only 1 minor node omitted or misidentified.	Nearly all edges correct; just 1 minor edge omitted or misidentified.	Structure largely matches; minor differences but no significant flaws.
7	Core nodes identified accurately, but minor omissions or misidentifications (1–2 non-critical nodes).	Most core edges correct; 1–2 non-critical edges missed or incorrect.	Clearly reasonable and coherent structure, minor noticeable flaws.
6	Most nodes correct, but clearly missing or misidentifying a few nodes.	Generally correct, but clearly missing or incorrectly identifying 1–2 important edges.	Generally reasonable but with clear structural errors or omissions.
5	Around half of the nodes correct; obvious omissions or errors.	Around half of the edges correct; obvious errors or omissions.	Obvious structural problems; overall logic still somewhat coherent.
4	Only a small portion of nodes correct; many omissions or errors.	Poorly identified; only a small portion of core edges correct.	Partially confusing; only some parts clearly reasonable.
3	Most nodes incorrect, only a few correct.	Mostly incorrect edges, only a few correct.	Mostly chaotic; few structurally reasonable elements.
2	Mostly incorrect; only one or two nodes correct by chance.	Only 1–2 edges correct.	Severe structural issues; only minor elements reasonable by chance.
1	Only one node identified correctly; all others wrong.	Almost entirely incorrect; only one edge correct by chance.	Nearly completely incorrect; minimal structural coherence by chance.
0	Completely incorrect; no correct nodes identified.	Completely incorrect; no correct edges identified.	Completely incorrect; no structural coherence at all.

Important Clarifications:

- You are to assign scores specifically to the model-generated causal graph, NOT to the Ground Truth causal graph.

- Your evaluation must be strictly based on comparing the model-generated causal graph with the provided Ground Truth causal graph and guided by the causal reasoning problem context, which clarifies the meaning of each node and edge.

- Evaluate separately along three independent dimensions:
• Node Accuracy(0–10 points)
• Edge Accuracy(0–10 points)
• Overall Structural Quality(0–10 points)

- Follow the detailed scoring criteria provided below, and briefly justify your rating for each dimension.
Detailed Scoring Criteria (0–10 points each dimension):

Node Accuracy:
- 10: All nodes perfectly identified, no errors or omissions.
- 9: All core nodes correctly identified; only minor discrepancies with non-critical nodes.
- 8: Nearly all nodes correctly identified; only 1 minor node omitted or misidentified.
- 7: Core nodes identified accurately, but minor omissions or misidentifications (1–2 non-critical nodes).
- 6: Most nodes correct, but clearly missing or misidentifying a few nodes.
- 5: Around half of the nodes correct; obvious omissions or errors.
- 4: Only a small portion of nodes correct; many omissions or errors.
- 3: Most nodes incorrect, only a few correct.
- 2: Mostly incorrect; only one or two nodes correct by chance.
- 1: Only one node identified correctly; all others wrong.
- 0: Completely incorrect; no correct nodes identified.

Edge Accuracy:
- 10: All edges (including directions) identified perfectly.
- 9: Nearly perfect; only one minor discrepancy on a non-critical edge.
- 8: Nearly all edges correct; just 1 minor edge omitted or misidentified.
- 7: Most core edges correct; 1–2 non-critical edges missed or incorrect.
- 6: Generally correct, but clearly missing or incorrectly identifying 1–2 important edges.
- 5: Around half of the edges correct; obvious errors or omissions.
- 4: Poorly identified; only a small portion of core edges correct.
- 3: Mostly incorrect edges, only a few correct.
- 2: Only 1–2 edges correct.
- 1: Almost entirely incorrect; only one edge correct by chance.
- 0: Completely incorrect; no correct edges identified.

Overall Structural Quality:
- 10: Structure perfectly matches the Ground Truth; fully reasonable.
- 9: Structure highly matches, minor irrelevant differences only.
- 8: Structure largely matches; minor differences but no significant flaws.
- 7: Clearly reasonable and coherent structure, minor noticeable flaws.
- 6: Generally reasonable but with clear structural errors or omissions.
- 5: Obvious structural problems; overall logic still somewhat coherent.
- 4: Partially confusing; only some parts clearly reasonable.
- 3: Mostly chaotic; few structurally reasonable elements.
- 2: Severe structural issues; only minor elements reasonable by chance.
- 1: Nearly completely incorrect; minimal structural coherence by chance.
- 0: Completely incorrect; no structural coherence at all.

Please strictly follow the JSON format below when returning your evaluation:

```
{
"Node_Accuracy": {"Score": (0–10), "Brief_Reasoning": "..."},
"Edge_Accuracy": {"Score": (0–10), "Brief_Reasoning": "..."},
"Overall_Structural_Quality": {"Score": (0–10), "Brief_Reasoning": "..."}
}
```

Now, proceed to your evaluation:

Causal Reasoning Problem Context:
{problem_context}

Ground Truth Causal Graph Description (Evaluation Standard):
{reasoning with ground truth causal graph}

Model-generated Causal Graph Description (Evaluation Target):
{LLM's output for the problem (including causal diagram)}

Example of LLM-based Evaluation of Causal Graph Quality To clearly illustrate the LLM-based evaluation procedure used in our experiments, we provide a detailed example. Below, we demonstrate step-by-step how we objectively assessed the quality of causal graphs generated by different training methods (BaseModel (Llama-3.1-8B), and CDCR-SFT-Enhanced).

Detailed Evaluation Example

(1) Causal Reasoning Problem Context:

Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships: Demand has a direct effect on supply and price. Yield per acre has a direct effect on supply. Supply has a direct effect on price. Demand is unobserved. The overall probability of increased supply is 60%. The probability of reduced supply and increased price is 25%. The probability of increased supply and increased price is 24%. Is the chance of increased price smaller when observing increased supply?

(2) Ground Truth Causal Graph (Evaluation Standard):

- **Nodes:** Let V2 = yield per acre; V1 = demand; X = supply; Y = price.
- **Edges:** V1 ->X, V2 ->X, V1 ->Y, X ->Y

(3) Model-generated Causal Graphs (Evaluation Targets):

BaseModel (Llama-3.1-8B) Raw Output:

To analyze the scenario, we need to construct a causal graph.

The causal graph for the scenario is as follows:

A (Demand) ->B (Supply)
B (Supply) ->C (Price)
B (Supply) ->D (Yield per acre)

From the given probabilities, we have the following conditional probability tables: [omit reasoning parts...]

CDCR-SFT-Enhanced (Llama-3.1-8B) Raw Output:

Alright, let me first review your input. Next, I will build a causal graph from the information provided, defining each node and clarifying how they interact. After that, I will detail the steps of causal inference, describing how I move from the causal graph to the final answer. To ensure clarity, I will begin by presenting the causal graph's structure, the meaning of each node, and their connections. Then, I will illustrate the inference process, leading up to the result.

Causal Graph:

First, here is the section on the causal graph nodes. For each node, I will list its ID, Name, and provide a brief description.

Nodes: Node ID: V1 Node Name: Demand Node Description: Unobserved variable representing demand. Directly influences both supply and price.

Node ID: Y Node Name: Price Node Description: Price level. Directly influenced by demand and supply.

Node ID: X Node Name: Supply Node Description: Supply level. Directly influenced by yield per acre and demand, and directly affects price.

Node ID: V2 Node Name: Yield per Acre Node Description: Yield per acre. Directly influences supply.

Next, I will explain how these nodes are linked in the causal graph, showing how information flows between them. I will go through each node in turn, indicating which nodes feed into it and which nodes it influences.

Edges: Node: V1 Inputs: N/A Outputs: ['X', 'Y']

Node: Y Inputs: ['V1', 'X'] Outputs: N/A

Node: X Inputs: ['V1', 'V2'] Outputs: ['Y']

Node: V2 Inputs: N/A Outputs: ['X']

[omit reasoning parts...]

(4) LLM-based Evaluation Results:
Below are the LLM-generated evaluation results, strictly adhering to our defined scoring criteria and JSON format described in Appendix B.4.

BaseModel Evaluation:


```
{
  "Node_Accuracy": {
    "Score": 6,
    "Brief_Reasoning": "The model-generated graph correctly identifies the core
        nodes (Demand, Supply, Price) but misidentifies the relationship with Yield
        per acre, which should be a direct effect on Supply rather than an edge from
         Supply to Yield."
  },
  "Edge_Accuracy": {
    "Score": 5,
    "Brief_Reasoning": "The model-generated graph has the correct edges for Demand
        to Supply and Supply to Price, but incorrectly includes an edge from Supply
        to Yield instead of Yield to Supply, leading to significant errors in edge
        identification."
  },
  "Overall_Structural_Quality": {
    "Score": 5,
    "Brief_Reasoning": "The structure is somewhat coherent with the core
        relationships but contains critical misrepresentations, particularly
        regarding the directionality of edges and the inclusion of Yield per acre,
        which affects the overall logical flow."
  }
}
```


CDCR-SFT-Enhanced Evaluation:


```
{
  "Node_Accuracy": {
    "Score": 10,
    "Brief_Reasoning": "All nodes (Demand, Price, Supply, Yield per Acre) are
        perfectly identified and accurately described."
  },
  "Edge_Accuracy": {
    "Score": 10,
    "Brief_Reasoning": "All edges are correctly identified with the correct
        directions, reflecting the causal relationships as per the Ground Truth."
  },
  "Overall_Structural_Quality": {
    "Score": 10,
    "Brief_Reasoning": "The structure of the model-generated causal graph perfectly
        matches the Ground Truth, with clear and coherent relationships among the
        nodes."
  }
}
```

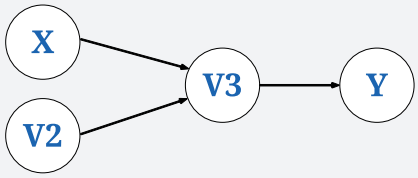

The above example explicitly illustrates the exact procedure and transparency of our evaluation methodology, starting from raw model outputs, extracting structured causal graphs, and finally obtaining standardized LLM-based scores.

B.5 Case Study Supplementary

To demonstrate the practical effectiveness of CDCR-SFT, we present a case illustration on the complete execution flow for a counterfactual reasoning query from WIQA that requires understanding complex causal mechanisms in Fig. 6. The selected query exemplifies the challenges that CDCR-SFT addresses: "*Process steps: An adult frog spawns eggs in water | These eggs hatch into tadpoles and continue to live in the water | The tadpoles grow developing external gills and a longer tail | The tadpole begins storing food in the tail | The tadpole develops hind legs and lives off food stored in the its tail | The front legs appear and the tadpoles tail shortens | The tadpole now looks like a young frog | The young frog continues to grow to adulthood. Question: "Will the absence of adult frogs laying eggs cause more frog population, less frog population, or have no effect?" Choice A: more; Choice B: less; Choice C: no effect*", which demands exact node identification, edge direction, and graph-based inference to trace effects without adding false links. The causal DAG construction phase breaks down the scenario into nodes and edges. The reasoning phase builds on the DAG to trace paths step by step following the reasoning goal. Finally, we got the answer.

Alright, let me first review your input. Next, I will build a causal graph from the information provided, defining each node and clarifying how they interact. After that, I will detail the steps of causal inference, describing how I move from the causal graph to the final answer. To ensure clarity, I will begin by presenting the causal graph's structure, the meaning of each node, and their connections. Then, I will illustrate the inference process, leading up to the result.
Causal Graph:
First, here is the section on the causal graph nodes. For each node, I will list its ID, Name, and provide a brief description.
Nodes:
Node ID: X\nNode Name: Adult Frog Reproduction\nNode Description: The ability of adult frogs to lay eggs, which directly influences the number of tadpoles and subsequently the frog population.
Node ID: V2\nNode Name: Tadpole Survival\nNode Description: The survival rate of tadpoles, affected by environmental factors and parental traits.
Node ID: V3\nNode Name: Tadpole Development\nNode Description: The developmental stage of tadpoles, influenced by their environment and parental traits through egg transmission.
Node ID: Y\nNode Name: Frog Population\nNode Description: The total number of frogs in the population, determined by survival and reproduction rates.

Next, I will explain how these nodes are linked in the causal graph, showing how information flows between them. I will go through each node in turn, indicating which nodes feed into it and which nodes it influences.
Edges:
Node: X\nInputs: N/A\nOutputs: ['V3']
\nNode: V2\nInputs: ['X']\nOutputs: ['V3']
Node: V3\nInputs: ['X', 'V2']\nOutputs: ['Y']
Node: Y\nInputs: ['V3']\nOutputs: N/A



Based on the current provided input information, the inference goal is that "*Determine whether the absence of adult frogs laying eggs causes an increase, decrease, or no effect in the frog population.*"

Given this goal, the following describes the **causal inference process.**

Causal Inference Process: The causal graph shows that adult frog reproduction (X) directly affects tadpole development (V3) and indirectly influences tadpole survival (V2). When X=0 (no egg-laying), V3=0 (no tadpoles), and V2=0 (no survivors). Since Y is fully determined by V3 and V2 (Y = V3 ∧ V2), Y=0. The absence of eggs eliminates all potential frogs, as tadpoles require eggs to develop. This causal pathway confirms that preventing adult frogs from laying eggs (X=0) reduces the population to zero, making the answer 'less.'
As a result of this causal inference process, I will reply with the **answer choice B**."

Figure 6: CDCR-SFT inference trace demonstrating the output integrating causal DAG, reasoning path, and answer for a counterfactual reasoning query.