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], Li 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,li.li02,yue.z}@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 4 LLMs across 8 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, Datasets — https://github.com/MrLYG/CDCR-SFT
Extended version — https://arxiv.org/abs/2508.12495

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

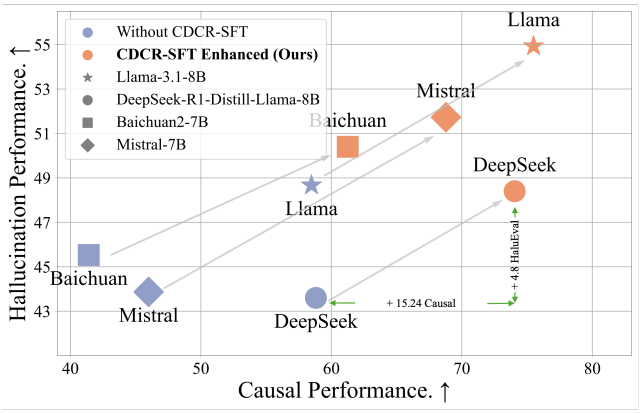

Figure 1: Average overall performance of our CDCR-SFT applied to four LLMs on the **causal reasoning** benchmarks (CLADDER and WIQA) and the **hallucination** benchmark (HaluEval). Orange symbols mark LLMs enhanced by CDCR-SFT, showing clear gains in causal reasoning and reduced hallucinations.

(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: "***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,

*These authors contributed equally.
†Corresponding authors.

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 methods 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-**D**AG and **R**easoning), 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

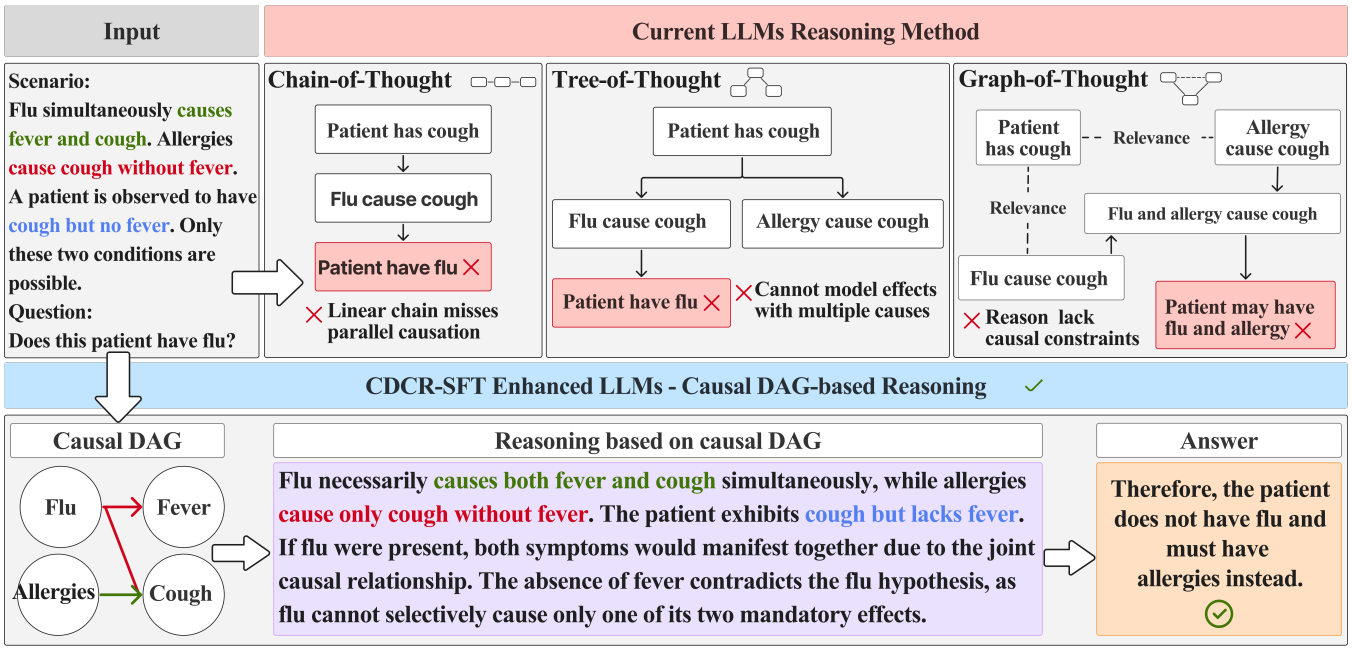

Figure 2: Comparison of reasoning approaches: Existing methods (CoT, ToT, GoT) operate at linguistic/semantic levels without explicit causal structure; Our CDCR-SFT constructs a variable-level causal DAG with directed edges representing causal relationships, enabling principled causal inference through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

Fig. 2. Existing structured reasoning methods, such as CoT, ToT, and GoT, generally produce reasoning paths at the linguistic 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	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 (Ours) 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 ID: V2;	Node ID: V3;	Node ID: Y;
Node Name: Flu;	Node Name: Allergy;	Node Name: Fever;	Node Name: Cough;
Node Description: Virus causing fever and cough simultaneously.	Node Description: Immune reaction causing cough only, without fever.	Node Description: Elevated body temperature, indicative of infection.	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). We construct CausalDR

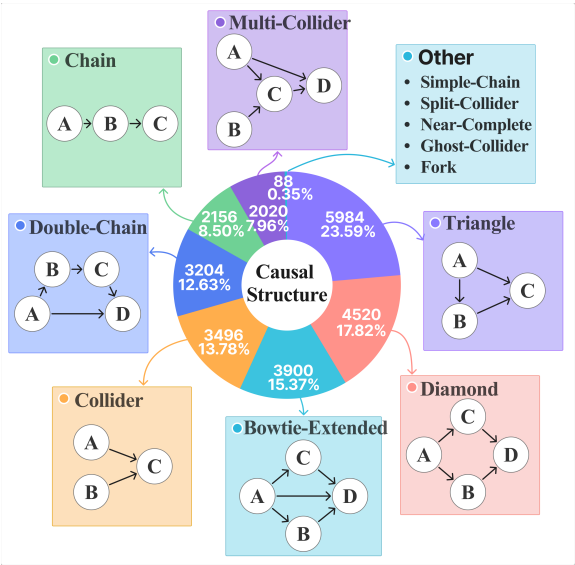

Figure 4: Proportional Distribution of 12 Canonical Causal DAG Structures in the CausalDR Dataset.

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. And Fig. 4 shows the distribution of CausalDR's 12 canonical DAGs, spanning chains, confounders, colliders, and multi-path interactions (e.g., Diamond, Bowtie-Extended).

Auxiliary Instruction following Data To avoid overspecialization on causal tasks and maintain the model's general linguistic capacity, we incorporate 10,000 randomly sampled Alpaca (Taori et al. 2023) instances into the supervised fine-tuning corpus alongside CausalDR.

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 4 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). We compare our CDCR-SFT method against 5 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. Additionally, DeepSeek-R1-Distill-Llama-8B, which is based on Llama-3.1-8B and fine-tuned on high-quality reasoning data (DeepSeek-AI 2025), serves as a supervised fine-tuning reasoning baseline.

Dataset	# Samples	Evaluation Focus
CLADDER	1,906	Causal reasoning; Causal DAG quality
WIQA	212	Causal reasoning
HaluEval	1,500	Hallucination

Table 2: Summary of datasets used in experiments.

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. **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), which changes within the text that test causal chain reconstruction, and out-of-paragraph (EXOGENOUS), which external changes assess the model's reasoning about external influences (more information 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.

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 4 LLMs on 2 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 methods 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

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.

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%. 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.

SFT-based comparison. DeepSeek-R1-Distill-Llama-8B serves as an SFT-trained reasoning baseline. Under the same Llama-3.1-8B, CDCR-SFT outperforms it on WIQA (55.66% vs. 51.42%) and HaluEval (54.93% vs. 43.60%), showing stronger causal reasoning and lower hallucination. These empirical findings directly support our core hypothesis: explicitly improving the causal reasoning capabilities of LLMs inherently mitigates 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 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

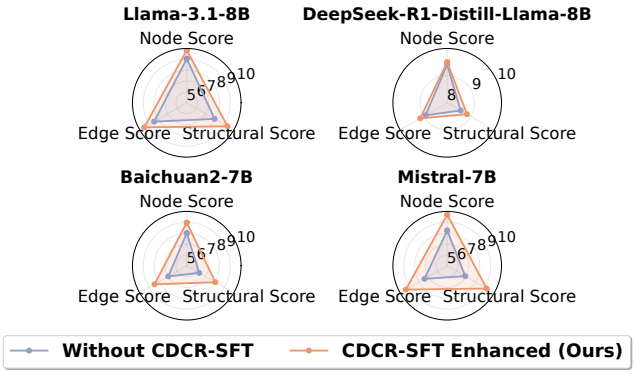

Figure 5: Comparison of causal DAG quality scores generated by pretrained LLMs versus those enhanced with CDCR-SFT, evaluated on the CLADDER dataset.

the causal DAG generated by pre-trained LLMs versus the causal DAG produced by LLMs enhanced with our CDCR-SFT approach, using the same prompt. 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 of over 2 points in both Edge Score and Structural Score.

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 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.

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.

5 Conclusion

We propose CDCR-SFT, which shifts LLM causal reasoning from sequential CoT or graph variants to causal DAG-based reasoning. It trains models to construct a causal DAG that encodes both causal directionality and conditional independencies, enabling them to perform structured inference over the graph. And we create the CausalDR dataset, which contains 25,368 validated samples, providing high-quality supervision for LLMs to learn explicit causal DAG construction and graph-based reasoning. Across 4 LLMs on CLADDER, WIQA, and HaluEval benchmarks, CDCR-SFT significantly improves causal reasoning, achieving SOTA accuracy of 95.33% on CLADDER (surpassing human performance of 94.8% for the first time) and reducing hallucinations on HaluEval by up to 11%. 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.

Acknowledgments

This work was partially supported by the National Science Foundation under Award No. 2428039, No. 2346158, and No. 2449280, Capital One Research Awards, and Amazon Research Awards. We also acknowledge the use of computational resources provided by the Advanced Cyberinfrastructure Coordination Ecosystem (Boerner et al. 2023): Services & Support (ACCESS) program, supported by NSF grants #2138259, #2138286, #2138307, #2137603, and #2138296. Specifically, this work used NCSA Delta GPU at the National Center for Supercomputing Applications (NCSA) through allocation CIS250073. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, Capital One, and Amazon.

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 AAAI conference on artificial intelligence*, volume 38, 17682–17690.

Boerner, T. J.; Deems, S.; Furlani, T. R.; Knuth, S. L.; and Towns, J. 2023. Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and experience in advanced research computing 2023: Computing for the common good*, 173–176.

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.