Causal-Audit: Explicit and Auditable Graph-based Reasoning via Target-Aware Causal Chain Construction

Su Lan, Xuefei Yin, Yanming Zhu, Alan Wee-Chung Liew*
Griffith University, Brisbane, Australia
{s.lan, x.yin, yanming.zhu, a.liew}@griffith.edu.au
*Corresponding author

Abstract

Causal and intervention-based question answering is fundamental to advancing large language models (LLMs) toward reasoning beyond surface-level correlations and understanding underlying causal mechanisms. However, existing LLM-based methods often rely on implicit language-level reasoning, resulting in opaque causal assumptions, unverifiable reasoning paths, and fragile predictions under complex interventions, particularly in context-free settings. In this paper, we propose an explicit and auditable causal reasoning framework for context-free intervention-based question answering. Our method formulates causal inference as structured reasoning over an explicit causal graph through four modular stages, rather than implicit end-to-end prediction. A key innovation is a target-aware causal graph construction strategy that treats the target variable as a core constraint during graph expansion, effectively suppressing irrelevant variables, spurious causal relations, and reasoning noise. We further introduce a path-level causal evidence aggregation mechanism that combines multiple causal paths while modeling both reinforcing and counteracting effects, enabling robust decision-making beyond single-chain reasoning. Extensive experiments on two causal-direction benchmarks and one medically grounded proxy benchmark demonstrate that our framework consistently outperforms existing LLM-based methods while providing interpretable and auditable reasoning traces.

1 Introduction

Causal and intervention-based question answering is widely regarded as a critical step in advancing large language models (LLMs) from powerful language processors toward reliable reasoning systems (Wu et al., 2024). Such questions ask how outcomes would change under hypothetical interventions rather than predicting what is typically observed, requiring models to reason about causal effects instead of exploiting statistical regularities (Jin et al., 2023; Zhou et al., 2024). Formally, it concerns the effect of actively manipulating a variable on a target outcome, often involving multiple interacting factors and indirect causal pathways.

Among different settings, *context-free* causal question answering, where models must infer causal effects without access to supporting passages or explicit background context, is particularly challenging (Pearl and Mackenzie, 2018). In this setting, answers cannot be derived via evidence extraction or surface-level alignment, but instead require the internal construction and evaluation of plausible causal mechanisms. In this work, we focus on this challenging yet fundamental setting of *context-free causal question answering*.

Recent advances in LLMs have led to encouraging performance gains on causal and counterfactual benchmarks, largely driven by end-to-end prediction or Chain-of-Thought style prompting (Wei et al., 2022; Kojima et al., 2022). However, a growing body of empirical studies suggests that such improvements often arise from pattern matching or shallow heuristics rather than genuine causal understanding (Miller et al., 2025). LLMs have been shown to struggle with intervention-level reasoning, multi-variable causal dependencies, and generalization beyond observed correlations, especially in settings that require reasoning about unseen or counterfactual scenarios (Jin et al., 2023; Wang, 2024; Zhou et al., 2024).

Several methods have been proposed to enhance LLMs for causal question answering, but most perform reasoning primarily at the language level, implicitly delegating causal inference to token- or semantic-level generation (Wei et al., 2022). Paradigms such as Chain-of-Thought (CoT) (Wei et al., 2022), Tree-of-Thought (Yao et al., 2023), and Graph-of-Thought (Besta et al., 2024) organize textual rationales without explicit variable-level causal structure or intervention semantics, mak-

Findings of the Association for Computational Linguistics: ACL 2026, pages 12484–12500

ing them vulnerable to spurious yet semantically related information. Although some methods introduce causal supervision during training (Li et al., 2025), inference typically collapses to direct question answering without explicit construction or auditing of causal graphs, leaving causal pathways, effect propagation, and interference from spurious variables unobservable and unverifiable (Gendron et al., 2024). Recent studies have consequently raised concerns about the reliability and faithfulness of such implicit reasoning, showing that generated rationales may not causally support final predictions (Paul et al., 2024; Chu et al., 2025). Moreover, methods that explicitly construct graphs from textual evidence rely on context-rich inputs and often fail in context-free settings, where relevant variables and relations cannot be reliably grounded (Tandon et al., 2019).

To address these limitations, we propose an explicit and auditable causal reasoning framework for intervention-based question answering. Rather than relying on implicit end-to-end prediction or free-form rationales, we formulate causal reasoning as structured inference over an explicit causal graph, exposing intermediate causal hypotheses and enabling inspection of the reasoning process via four modular stages. A key contribution is a *target-aware* causal graph construction strategy that treats the target variable as a core constraint during graph expansion, effectively suppressing irrelevant variables, spurious relations, and reasoning noise. Moreover, we introduce a path-level causal evidence aggregation mechanism that combines multiple causal paths and models both reinforcing and counteracting effects, moving beyond single-chain reasoning. These designs transform LLMs from implicit end-to-end "reasoners" into constrained "causal evaluators", offering a new paradigm that integrates LLMs with symbolic causal reasoning for reliable intervention-based question answering. Extensive experiments on three datasets demonstrate consistent performance gains over existing LLM-based methods while providing interpretable and auditable reasoning traces.

The main contributions are summarized below:

- We formulate context-free intervention-based question answering as an explicit and auditable causal reasoning problem, moving beyond implicit language-level inference and providing a structured framework for reliable intervention reasoning with LLMs.
- We propose a target-aware causal graph construction that explicitly treats the target variable as a core constraint during graph expansion, effectively suppressing irrelevant variables, spurious causal relations, and reasoning noise in traditional graph-based reasoning.
- We introduce a path-level causal evidence aggregation mechanism that validates and combines multiple causal paths while modeling both reinforcing and counteracting effects, enabling robust decision making beyond single-chain or end-to-end methods.
- Extensive experiments on three benchmarks demonstrate that our framework consistently outperforms existing LLM-based reasoning methods while providing interpretable and auditable causal reasoning traces.

2 Related Work

2.1 Direct LLM

Recent benchmarks such as CausalBench (Wang, 2024; Zhou et al., 2024), CLadder (Jin et al., 2023), and CausalProbe (Chi et al., 2024) show that, despite strong surface-level performance, LLMs often rely on shallow correlations rather than genuine causal understanding, especially under complex interventions. Counterfactual and interventional reasoning further emerges inconsistently through in-context learning and remains fragile to prompt design and distribution shifts (Miller et al., 2025). These reveal a persistent gap between correlation-based inference and causal structure learning.

2.2 Chain-of-Thought Reasoning

CoT prompting improves multi-step reasoning (Wei et al., 2022; Kojima et al., 2022), with self-consistency further enhancing robustness (Wang et al., 2023). However, CoT-based methods rely on unstructured natural language rationales, lack causal and structural constraints, and are often not causally faithful to final predictions (Paul et al., 2024; Chu et al., 2025). Graph-based prompting and neuro-symbolic methods introduce structural scaffolding (Besta et al., 2024; Fang et al., 2024) but do not explicitly encode causal semantics. In contrast, our work grounds graph construction in causal relations and validates reasoning at the path level.

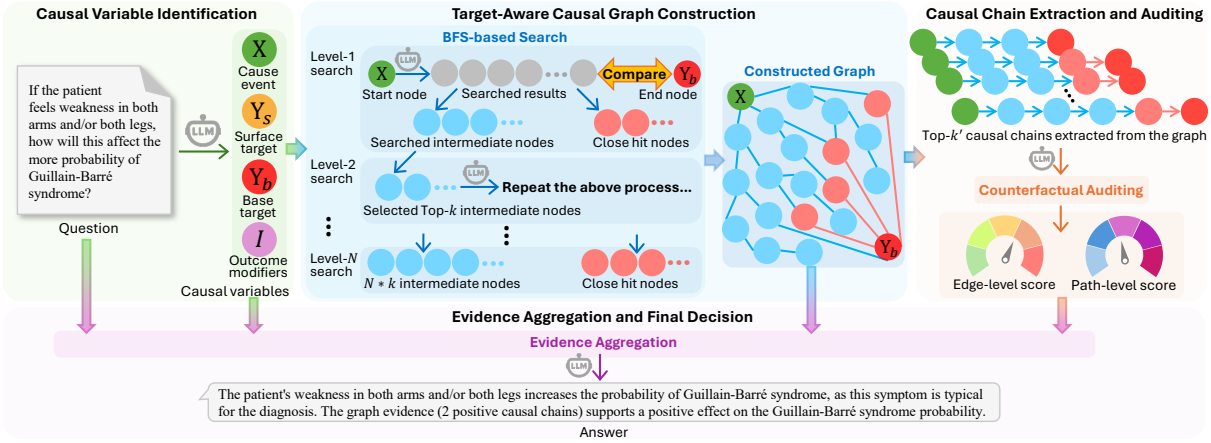

Figure 1: Overview of the proposed explicit causal reasoning framework for context-free intervention-based question answering. The framework extracts structured causal variables from the question, constructs a target-aware causal graph via iterative search, audits candidate causal chains through counterfactual validation, and aggregates evidence from multiple stages to produce an interpretable intervention decision. Detailed algorithmic descriptions are provided in Appendix A.

2.3 Causal Graph-based Reasoning

Integrating causal graphs and causal inference into LLM reasoning has emerged as a promising direction for improving faithfulness and interpretability. Existing methods incorporate causal structures (Wang et al., 2024) or employ causal objectives and structured supervision (Dong et al., 2025; Li et al., 2025), yet intervention-based analyses reveal frequent violations of causal faithfulness (Paul et al., 2024). Moreover, most methods rely on predefined or externally provided graphs. In contrast, our framework dynamically constructs target-aware causal graphs and performs path-level validation and evidence aggregation during inference.

2.4 Knowledge Graph-Augmented Reasoning

Another line of work augments LLM reasoning with external knowledge graphs (KGs). Path-based methods select informative KG paths to support multi-hop reasoning (Liu et al., 2024; Tan et al., 2025). Recent benchmarks evaluate LLMs' use of KG structure (Markowitz et al., 2025), with causal-aware graph augmentation further explored (Luo et al., 2025). However, these methods treat paths primarily as evidence and do not explicitly model causal direction or competing effects. Our framework instead models signed causal influence and aggregates conflicting evidence during inference.

3 Methodology

Overview. Given a context-free intervention-based causal question, the goal is to determine the directional effect of an intervention variable on a target outcome. To this end, we propose an explicit causal reasoning framework that constructs, audits, and aggregates causal evidence in a structured and interpretable manner. As illustrated in Fig. 1, the framework consists of four sequential stages. First, structured query extraction identifies the intervention variable, target variables, and outcome modifiers from the input question. Second, a target-aware causal graph is constructed through iterative expansion, where graph growth is explicitly constrained by relevance to the target variable. Third, candidate causal chains connecting the intervention and target are extracted from the graph and audited to ensure causal consistency under hypothetical interventions. Finally, validated causal evidence from the first three stages is aggregated to produce the final decision on the direction of the intervention effect. This staged formulation enables systematic control of the reasoning process, improves interpretability through explicit causal structures, and allows only causally validated evidence to contribute to the final prediction.

3.1 Causal Variable Identification

Given a natural language question q, we employ an LLM-based extractor Φ_{extract} to produce a structured query

$$\mathcal{Q} = (X, Y_s, Y_b, I). \tag{1}$$

Here, X denotes the *intervention variable* representing the manipulated cause event; Y_s is the *surface target*, i.e., the surface-level outcome expres-

sion in the question; Y_b is the *base target variable*, a canonical scientific noun phrase used for causal reasoning; and I contains *outcome modifiers*, including the directional indicator $D \in \{\text{MORE}, \text{LESS}\}$ and a negation flag. We distinguish Y_s from Y_b so that graph construction and causal reasoning are performed on the normalized variable pair (X, Y_b), while the final prediction is mapped back to the original question semantics using I. This separation decouples causal reasoning from surface linguistic variation, enabling more stable graph construction and consistent reasoning across semantically equivalent questions.

3.2 Target-Aware Causal Graph Construction

We propose to construct a directed causal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ rooted at the intervention variable X, with the objective of identifying causal mechanisms leading to the base target variable Y_b.

BFS Expansion with Negative Constraints. We adopt a breadth-first search (BFS) expansion strategy (Kozen, 1992) that iteratively explores candidate causal effects layer by layer to construct the graph. This process incrementally grows the graph outward from X, allowing the framework to capture multi-hop causal mechanisms while maintaining explicit control over the search depth and expansion scope. We initialize the frontier $\mathcal{F}_0 = \{X\}$ and the visited set $\mathcal{V}_{\text{vis}} = \{X\}$, where the frontier represents the set of nodes to be expanded at the current step, and the visited set records all nodes observed so far. At each expansion step t, we collect a pool of LLM-proposed candidate causal triples for the next layer, denoted by $\mathcal{C}_{t+1}$:

$$\mathcal{C}_{t+1} = \bigcup_{u \in \mathcal{F}_t} \Big\{ (u, r, v) \;\Big|\; (u, r, v) \sim \Phi_{\text{expand}}(u \mid \text{Avoid} = \mathcal{V}_{\text{vis}}) \Big\}. \tag{2}$$

Here, (u, r, v) represents a directed causal relation $u \rightarrow v$ with effect polarity $r \in \{\text{INC}, \text{DEC}\}$, where INC denotes $u \uparrow \Rightarrow v \uparrow$ and DEC denotes $u \uparrow \Rightarrow v \downarrow$. The $\sim$ denotes LLM generation or decoding. Injecting the visited set as a negative constraint discourages revisiting previously explored nodes, reducing redundant loops and cyclic reasoning. This design enforces acyclic graph growth and prevents uncontrolled expansion, keeping the search focused on plausible causal mechanisms.

Fine-Grained Variable Alignment (FGVA). To determine whether a generated node v is relevant to the base target Y_b, we introduce a fine-grained variable alignment function $\Phi_{\text{align}}(v, Y_b)$ that evaluates three dimensions: (i) *Entity alignment* $S_E \in \{\text{Exact}, \text{Partial}, \text{None}\}$; (ii) *Quantity alignment* $S_Q \in \{\text{Exact}, \text{Subset}, \text{Agg}, \text{None}\}$; and (iii) *State alignment* $S_S \in \{\text{Match}, \text{Conflict}, \text{None}\}$. Intuitively, S_E assesses whether the core entity or topic matches (or partially overlaps), S_Q checks whether v and Y_b refer to the same granularity (e.g., exact vs. subset/aggregate), and S_S checks whether the described state is logically compatible. By jointly considering these dimensions, FGVA enables relevance assessment beyond surface string similarity, allowing the framework to distinguish target-equivalent variables, near-target proxies, and intermediate bridging concepts.

Based on the alignment results, we classify each node v into a relevance class:

$$Cls(v) = \begin{cases} \text{EXACT}, & S_E^{ex} \wedge S_Q^{ex} \wedge S_S^{m}, \\ \text{CLOSEHIT}, & S_E^{ex} \wedge S_Q^{sa} \wedge \neg S_S^{c}, \\ \text{BRIDGE}, & S_E^{pa} \wedge \neg S_S^{c}, \\ \text{NONE}, & \text{otherwise}. \end{cases} \tag{3}$$

where $S_E^{ex} \triangleq [S_E = \text{Exact}]$, $S_E^{pa} \triangleq [S_E = \text{Partial}]$, $S_Q^{ex} \triangleq [S_Q = \text{Exact}]$, $S_Q^{sa} \triangleq [S_Q \in \{\text{Subset}, \text{Agg}\}]$, $S_S^{m} \triangleq [S_S = \text{Match}]$, and $S_S^{c} \triangleq [S_S = \text{Conflict}]$. The assigned class $Cls(v)$ determines how the node is used during graph construction. Consequently, we maintain three auxiliary sets: $\mathcal{N}_{\text{exact}} = \{v \mid Cls(v) = \text{EXACT}\}$, $\mathcal{N}_{\text{close}} = \{v \mid Cls(v) = \text{CLOSEHIT}\}$, and $\mathcal{N}_{\text{bridge}} = \{v \mid Cls(v) = \text{BRIDGE}\}$, while nodes with $Cls(v) = \text{NONE}$ are discarded and not expanded further. EXACT nodes are treated as target-equivalent and serve as terminal endpoints for causal-chain extraction. CLOSEHIT nodes act as semantic landing points near Y_b and support subsequent alignment toward the target. BRIDGE nodes are retained only as intermediate candidates and may continue to be expanded to reach CLOSEHIT or EXACT nodes, helping reduce semantic drift while preserving recall.

Target-Aware Pruning and Frontier Update. To control the exponential growth of the search space during graph expansion, we introduce a target-aware pruning strategy at each BFS step. Unconstrained expansion rapidly introduces loosely related or generic nodes, diluting causal signals and increasing the risk of spurious reasoning paths (Sui et al., 2022). Instead of heuristic string matching or static embedding similarity, we propose to use an LLM-based relevance ranking function $\Phi_{\text{rank}}(v, Y_b)$

that explicitly conditions on the base target variable Y_b. This ranking prioritizes nodes that are not only semantically related, but also causally promising for reaching the target domain.

At each expansion step, only the top-K ranked nodes are retained as the next frontier:

$$\mathcal{F}_{t+1} = \text{Top-}K(\mathcal{C}_{t+1}), \;\; \mathcal{V}_{vis} \leftarrow \mathcal{V}_{vis} \cup \mathcal{F}_{t+1}. \quad (4)$$

This goal-directed pruning focuses the expansion on plausible causal mechanisms while maintaining coverage of relevant effects. The process terminates when Y_b is reached or when a maximum depth D is exceeded.

Bridging Close-Hit Nodes. When the BFS expansion does not explicitly reach Y_b, we perform a bridge step to connect near-target nodes in $\mathcal{N}_{close}$. For each $u \in \mathcal{N}_{close}$, a verifier Ψ_{bridge} evaluates whether a direct causal relation is plausible; if so, a directed edge $(u, r_{\text{bridge}}, Y_b)$ is added to $\mathcal{E}$.

3.3 Causal Chain Extraction and Auditing

To derive causally valid explanations for intervention effects, we extract and audit causal chains connecting the intervention variable X to the base target Y_b from the constructed graph $\mathcal{G}$. We enumerate simple paths $\mathcal{P} = \{p_1, \ldots, p_m\}$ connecting X and Y_b, rank them using a structural cost function that penalizes excessive length, semantic drift, and generic nodes, and retain the top-K' paths for a two-stage causal audit.

Premise Consistency Check. A valid causal explanation must not contradict the intervention premise. For a path p, we define a premise validity indicator

$$V_{\text{prem}}(p) = \prod_{v \in p \setminus \{X, Y_b\}} \mathbb{I}[\neg C(v, X)]\,. \quad (5)$$

where $C(v, X)$ indicates if introducing an intermediate node v would violate the intervention premise on X. Paths with $V_{\text{prem}}(p) = 0$ are discarded.

Counterfactual Edge Verification. Each remaining path is further audited via counterfactual probing. For each edge $e = (u, v)$, an LLM estimates whether removing or reversing u induces a change in v, producing a edge-level confidence $V_{\text{cf}}(e) \in [0, 1]$. The final path-level audit score is defined as

$$S_{\text{audit}}(p) = V_{\text{prem}}(p) \cdot \prod_{e \in p} V_{\text{cf}}(e). \quad (6)$$

Paths with multiplicative audit score $S_{\text{audit}}(p)$ less than a predefined threshold τ_{audit} are removed.

3.4 Evidence Aggregation and Final Decision

Path-Level Effect and Weighting. Each causal relation r is assigned a signed effect $\sigma(r) \in \{+1, -1\}$. The net causal effect of a path p is computed as

$$\Delta(p) = \prod_{e \in p} \sigma(r_e), \quad (7)$$

reflecting whether the intervention increases or decreases the target outcome. To downweight weakly supported or implicit reasoning, each path is assigned a weight

$$w(p) = S_{\text{audit}}(p) \cdot \exp(-\gamma N_{\text{br}}(p)), \quad (8)$$

where $N_{\text{br}}(p)$ denotes the number of bridge edges.

Graph-Level Evidence Aggregation. We aggregate signed evidence across all audited paths $\mathcal{P}^*$. Let W^+ and W^- denote the total positive and negative evidence mass, respectively. The resulting graph-level confidence is defined as

$$C_{\text{graph}} = \frac{|W^+ - W^-|}{W^+ + W^- + \epsilon} \cdot \log(1 + |\mathcal{P}^*|)\,, \quad (9)$$

which jointly captures evidence dominance and support size. The corresponding graph decision $E_{\text{graph}} \in \{\text{MORE}, \text{LESS}\}$ is determined by the sign of $W^+ - W^-$.

Conflict Resolution with Evidence-Conditioned LLM. When causal evidence is weak or conflicting, we defer to an LLM prediction conditioned on the audited graph evidence. We serialize $\mathcal{P}^*$ into a compact evidence packet $\mathcal{Z}(\mathcal{P}^*)$, containing for each path its signed effect $\Delta(p)$, weight $w(p)$, and a short edge-sequence representation.[1] Formally,

$$E_{\text{LLM}} = f_\theta\big(q, \mathcal{Z}(\mathcal{P}^*)\big) \in \{\text{MORE}, \text{LESS}\}. \quad (10)$$

The final prediction is obtained by resolving E_{graph} and E_{LLM} using the graph confidence:

$$\hat{y} = \begin{cases} E_{\text{graph}}, & C_{\text{graph}} > \tau \text{ or } E_{\text{graph}} = E_{\text{LLM}}, \\ E_{\text{LLM}}, & \text{otherwise.} \end{cases} \quad (11)$$

Finally, $\hat{y}$ is mapped to the surface-level answer using the modifier set I.

[1] Deterministic decoding is used.

4 Experiments

4.1 Datasets, Task, and Evaluation Metric

Datasets and Task. We evaluate our method on three benchmarks spanning diverse domains and knowledge requirements: (i) DDXPlus-CausalEffect (medical); (ii) the directional subset of WIQA; and (iii) a general-domain dataset derived from CauseNet. All datasets are cast into a unified *context-free causal direction prediction* setting. Each instance specifies an intervention on a cause variable or event X and asks whether the target variable or event Y becomes `more` or `less` likely under the intervention. Following prior protocols on WIQA, we exclude NoEff instances and retain the original two-choice answer space. For DDXPlus-CausalEffect and the CauseNet-derived dataset, questions are rendered using a WIQA-style template with the same label set $\{\texttt{more}, \texttt{less}\}$.

DDXPlus-CausalEffect requires an important qualification. Its labels are association-based directional proxies rather than formally identified causal effects. We include it as a medically grounded proxy benchmark because the underlying DDXPlus resource is rooted in curated clinical knowledge and rule-based diagnostic structure rather than open-domain surface co-occurrence. We therefore interpret results on DDXPlus as complementary evidence under noisy proxy supervision, while WIQA and the CauseNet-derived benchmark remain our primary evidence for context-free causal direction prediction. Dataset construction details and statistics are provided in Appendix B. **Evaluation Metric.** Following prior works, we report *Accuracy* as the evaluation metric. Since the task is a binary directional classification with a fixed label space and follows standard evaluation protocols on the directional subset, Accuracy provides a clear and comparable measure across backbones and methods. Label distributions are reported in Appendix B for completeness.

4.2 Experimental Setup

Decoding and Reproducibility. All LLM calls use deterministic decoding (temperature $= 0$) with fixed decoding hyperparameters (e.g., `top_p` and `max_tokens`). When supported by the backend, we additionally fix the random seed to improve run-to-run consistency. Outputs are strictly constrained via JSON/schema validation; invalid responses are retried using a predefined template fallback.

Prompt Format. All methods adopt structured prompts with a fixed output schema, where the final prediction is constrained to $\{\texttt{more}, \texttt{less}\}$ and optional fields are used for intermediate reasoning or evidence. Full prompts, schemas, and example instances are provided in Appendix C.

Graph Construction and Auditing. Graph expansion is performed with bounded depth $D = 4$ and Top-$K = 2$ frontier selection per hop. For each frontier node, up to $R = 3$ candidate relations are generated, and simple paths from X to the (base) target are enumerated up to length $L = 6$. Candidate paths are filtered using an audit threshold $\tau_{\text{audit}} = 0.6$ before evidence aggregation. All hyperparameters are selected on a held-out development split (or via the grid search in Sec. 4.5) and fixed across all test runs.

4.3 Backbone Models and Compared Methods

Backbone Models We evaluate our method on multiple instruction-tuned LLM backbones from different model families at comparable scales, including Llama-3.1-8B, Mistral-7B, and Ministral-3-8B. To isolate the effect of the proposed reasoning framework, we keep the dataset, instance representation (explicit (X, Y)), prompt format, input/output schema, and decoding hyperparameters fixed for each backbone, and vary only the backbone model. This allows us to assess whether performance gains generalize across model families rather than relying on a specific backbone.

Compared Methods We compare our method with representative LLM-based reasoning baselines under a unified evaluation pipeline. Specifically, we include (i) *Direct LLM*, which directly predicts the directional label from (X, Y) without intermediate structure, serving as a parametric-knowledge baseline; (ii) *Chain-of-Thought* (COT) prompting (Wei et al., 2022), which elicits free-form step-by-step rationales before prediction; (iii) structured reasoning methods such as *Tree-of-Thought* (ToT) (Yao et al., 2023) and *Graph-of-Thought* (GoT) (Besta et al., 2024), which explore multiple reasoning paths via tree- or graph-structured deliberation without explicit causal modeling or verification; and (iv) CDCR-SFT (Li et al., 2025), which enhances causal reasoning by supervised LoRA fine-tuning on CausalDR, encouraging LLMs to explicitly construct causal DAGs before making predictions. All baselines use the same final answer constraint $\{\texttt{more}, \texttt{less}\}$ and identical decoding settings unless otherwise specified.

Datasets	Backbone Models	Direct LLM	CoT	GoT	ToT	Ours
DDXPlus-CausalEffect	Llama-3.1-8B	66.00	59.50	57.50	58.00	**67.00**
	Mistral-7B	43.50	40.00	46.00	38.00	**74.50**
	Ministral-3:8b	42.50	52.50	43.00	60.5	**81.50**
WIQA (direction subset)	Llama-3.1-8B	56.13	50.00	51.42	50.94	**67.92**
	Mistral-7B	31.13	37.74	41.51	37.26	**65.09**
	Ministral-3:8b	44.34	54.72	56.13	51.89	**60.38**
CauseNet-derived (direction)	Llama-3.1-8B	68.00	77.00	73.00	77.00	**79.00**
	Mistral-7B	72.00	81.00	67.00	50.00	**87.00**
	Ministral-3:8b	66.00	69.00	71.00	63.00	**79.00**

Table 1: Accuracy (%) on context-free causal direction prediction across three datasets and multiple LLM backbones. Free-form reasoning methods (CoT, GoT, ToT) exhibit high variance across backbones. Ours consistently achieves the best performance across backbones and domains, demonstrating robust gains from explicit causal graph construction, auditing, and evidence aggregation. Accuracy is reported under a best-of-multiple-attempts setting.

Backbone Models	CDCR-SFT	Ours	Δ
Llama-3.1-8B	55.66	**67.92**	+12.26
Mistral-7B	44.81	**65.09**	+20.28

Table 2: Accuracy (%) on WIQA (direction subset) compared with CDCR-SFT, a training-based method that learns causal DAG construction via supervised fine-tuning. Our method achieves substantially higher accuracy on both backbones without task-specific fine-tuning, highlighting the effectiveness of explicit and auditable causal modeling at inference time.

4.4 Main Results

Table 1 reports accuracy (%) for context-free causal direction prediction across datasets and backbones. Overall, **ours** consistently outperforms all baselines. We highlight three recurring patterns.

(1) Reliability under medically grounded proxy supervision. On DDXPlus-CausalEffect, which involves specialized medical knowledge and non-trivial causal dependencies, unconstrained reasoning baselines (CoT, ToT, GoT) are often unstable and can underperform Direct LLM prediction. This suggests that free-form rationales tend to introduce unsupported or domain-invalid mechanisms in expert domains. In contrast, **ours** consistently improves performance across all backbones, with particularly large gains on Mistral-7B (+28.5) and Ministral (+21.0). These results indicate that our proposed explicit construction and auditing of causal chains is crucial for reliable intervention reasoning in knowledge-intensive settings. Because DDXPlus uses association-based directional proxy labels rather than formally identified causal effects, we interpret these gains as evidence that explicit construction and auditing remain robust under noisy medically grounded supervision, rather than as a claim of causal identification on this benchmark.

(2) Strong gains on complex multi-hop causal reasoning. On WIQA, which requires multi-hop causal reasoning under interventions, **ours** yields substantial improvements over the strongest baseline across all backbones (+11.8, +23.6, and +4.3 points for Llama-3.1-8B, Mistral-7B, and Ministral, respectively). Notably, existing reasoning paradigms exhibit high variance across backbones, whereas **ours** consistently achieves the best performance. This suggests that our proposed explicit construction and auditing of causal structures is more effective than relying on implicit deliberation alone, particularly for long-horizon causal inference. By converting global reasoning into target-aware expansion and local edge verification, our framework provides a reliable inductive bias that generalizes across model families. To assess robustness to evaluation size, we additionally evaluate on a larger WIQA slice with 650 instances under the same task and labeling protocol. The gain remains robust: our method achieves 67.54% accuracy versus 53.69% for the Direct baseline. Detailed results are reported in Appendix D.1.

(3) Robustness in open-domain causal reasoning. On the open-domain CauseNet-derived benchmark, where spurious co-occurrence is common, purely deliberative or search-based reasoning methods exhibit high variance across backbones. For example, ToT performs competitively on Llama-3.1-8B but collapses on Mistral-7B. In contrast, **ours** consistently achieves strong performance across all back-

Hyperparameters			Performance		
DEPTH (D)	WIDTH (R)	PATHLEN (L)	Acc (%)	Exo-Acc (%)	In-Acc (%)
2	3	4	64.15	67.92	60.38
	3	6	66.04	70.75	61.32
	5	4	62.26	66.98	57.55
	5	6	63.21	69.81	56.60
4	3	4	62.74	66.98	58.49
	3	6	**67.92**	70.75	**65.09**
	5	4	66.04	70.75	61.32
	5	6	65.09	68.87	61.32
6	3	4	63.21	67.92	58.49
	3	6	65.57	67.92	63.21
	5	4	67.45	**71.70**	63.21
	5	6	66.98	69.81	64.15

Table 3: Hyperparameter sensitivity analysis on WIQA (direction subset) with the Llama-3.1-8B backbone.

bones, outperforming the best baseline on Mistral-7B (+6.0) and Ministral (+8.0). This robustness highlights the benefit of our proposed target-aware pruning and evidence aggregation in suppressing spurious causal paths and maintaining reliable inference in noisy, open-domain settings.

Comparison with causal-structure fine-tuning. To further contextualize our WIQA results, we compare against CDCR-SFT, a recent method that improves causal reasoning by supervised LoRA fine-tuning on the CausalDR dataset to explicitly learn causal DAG construction during training. As shown in Table 2, **ours** substantially outperforms CDCR-SFT on both backbones, with gains of +12.26 on Llama-3.1-8B and +20.28 on Mistral-7B. Notably, our approach achieves these improvements without any task-specific fine-tuning, demonstrating that explicit and auditable causal modeling at inference time can be more effective and generalizable than embedding causal structure implicitly into model parameters.

4.5 Hyperparameter Sensitivity Analysis

We analyze the sensitivity of our target-aware causal graph construction to three key hyperparameters: maximum expansion depth ($D \in \{2, 4, 6\}$), branching width ($R \in \{3, 5\}$), and maximum path length ($L \in \{4, 6\}$). Results are reported in Table 3 on WIQA with the Llama-3.1-8B backbone.

Effect of Expansion Depth. Performance improves when increasing the depth from $D = 2$ to a moderate range ($D = 4$), reflecting the necessity of multi-hop causal mechanisms for WIQA-style intervention questions. However, further increasing depth ($D = 6$) yields diminishing or unstable gains, suggesting that overly deep expansion introduces semantically weak or noisy intermediate variables. This trend supports our design choice of bounded expansion with auditing, which prioritizes causal relevance over exhaustive search.

Interaction Between Branching and Path Length. We observe a clear trade-off between branching width (R) and allowed path length (L). Moderate branching with sufficient path length (e.g., $R = 3, L = 6$) consistently yields strong performance, suggesting that accurate intervention reasoning relies on precise multi-hop causal chains rather than broad, shallow exploration. In contrast, aggressive early branching often degrades accuracy, particularly for shorter paths, due to the introduction of spurious or weakly related variables. These trends directly motivate our target-aware pruning and path-level auditing: by constraining early expansion and validating longer chains, our framework favors causally coherent multi-hop mechanisms while suppressing noisy alternatives.

Exogenous vs. Internal Interventions. Exo-Acc and In-Acc measure accuracy when the intervention variable appears outside or inside the paragraph, respectively. Higher In-Acc reflects better alignment and reduced semantic drift, while gains on Exo-Acc indicate stronger context-free generalization. Exo-Acc benefits from larger expansion depth due to longer mechanism chains, whereas In-Acc typically peaks at moderate depth. Their complementary depth preferences validate our bounded, target-aware expansion design.

4.6 Component Ablations

We ablate three core components on WIQA with the Llama-3.1-8B backbone: target-aware pruning (TAP), fine-grained variable alignment (FGVA), and counterfactual auditing (CFA). In addition to accuracy, we report *Path Reach*, the percentage of instances for which the corresponding pipeline retains at least one path from the intervention variable X to the base target Y_b. Table 4 shows that all three components contribute materially to final performance.

FGVA is particularly important: replacing it with similarity-based alignment sharply reduces both accuracy and Path Reach, indicating that semantically loose matching breaks graph connectivity and introduces spurious target links. CFA improves accuracy with only a minor reduction in Path Reach, suggesting that auditing mainly removes noisy links rather than broadly collapsing

Component	Setting	Acc.	Path Reach
TAP	Target-aware pruning	**67.92**	96.23
	Random pruning	64.60	**97.17**
FGVA	Fine-grained alignment	**67.92**	**96.23**
	Simple alignment	55.66	14.62
CFA	Counterfactual auditing	**67.92**	96.23
	No auditing	60.38	**97.28**

Table 4: Ablations on WIQA. Path Reach is the percentage of instances for which the corresponding pipeline retains at least one usable path from X to Y_b.

the graph. TAP improves accuracy over random pruning by trading a small amount of coverage for higher target relevance. Additional analyses of τ_{audit}, bridge penalty γ, audit robustness, efficiency, and failure cases are reported in Appendix D.

4.7 Case Study and Interpretability

We present a representative case study to illustrate how explicit causal modeling improves both reliability and interpretability. While standard CoT reasoning produces free-form narratives whose causal validity is difficult to assess, our framework externalizes reasoning as explicit causal graphs, audited chains, and quantified evidence. In a DDXPlus-CausalEffect example predicting the effect on *Chagas probability*, CoT incorrectly outputs `more` by invoking a generic prior ("early-stage disease may be asymptomatic") that is not causally grounded in the intervention. In contrast, our method constructs multiple candidate causal chains from the intervention variable, filters invalid mechanisms via counterfactual edge verification, and aggregates only premise-consistent evidence. This process yields the correct `less` prediction and exposes the specific causal paths supporting the decision. We provide the full constructed graph, audited chains, and per-edge counterfactual scores in Appendix E.

5 Conclusion

We propose an explicit, auditable causal reasoning framework for context-free intervention question answering. It formulates inference as structured reasoning via target-aware causal graph construction, path-level auditing, and evidence aggregation. By constraining variable expansion and verifying causal relations via counterfactual probing, it transparently resolves competing causal mechanisms. This design offers a clean paradigm for integrating LLMs with symbolic causal reasoning, transforming LLMs from end-to-end reasoners into constrained causal evaluators. Extensive experiments across a medically grounded proxy benchmark, a commonsense benchmark, and an open-domain benchmark demonstrate consistent and substantial improvements over existing methods, underscoring the importance of explicit causal structure and verification for reliable causal reasoning.

Limitations

Inference Complexity. The framework requires multiple LLM calls for graph expansion, alignment, and auditing, making it substantially more expensive than single-pass prompting. Although much of this cost is parallelizable across independent expansions and candidate paths, the method still involves a deliberate trade-off between reasoning reliability and efficiency. We report detailed timing and token statistics in Appendix D.3.

Boundary of Parametric Knowledge. The context-free setting does not assume that the backbone has memorized all domain-specific knowledge. Rather, it removes external retrieval, so all methods in this setting necessarily rely on parametric knowledge. Our framework can still succeed when the intervention can be mediated through general mechanisms already represented in the backbone, but it can fail when correct prediction depends on newly emerging medical findings, rare terminology, or highly specific domain facts that are not internalized by the model. In such cases, sparse or conflicting graph evidence should be interpreted as uncertainty rather than strong causal support.

Potential Risk of Misuse. The framework produces directional causal predictions with structured reasoning traces, but does not perform formal causal identification based on do-calculus. In sensitive domains such as healthcare, the outputs should therefore be interpreted as supportive reasoning signals rather than definitive causal or clinical conclusions.

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.

Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga

Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. 2022. Causalqa: A benchmark for causal question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3296–3308.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2024. Unveiling causal reasoning in large language models: Reality or mirage? volume 37, pages 96640–96670.

Zheng Chu, Jingchang Chen, Zhongjie Wang, Guo Tang, Qianglong Chen, Ming Liu, and Bing Qin. 2025. Towards faithful multi-step reasoning through fine-grained causal-aware attribution reasoning distillation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2291–2315.

Juncheng Dong, Yiling Liu, Ahmed Aloui, Vahid Tarokh, and David Carlson. 2025. Care: Turning llms into causal reasoning expert. In *First Workshop on Foundations of Reasoning in Language Models*.

Meng Fang, Shilong Deng, Yudi Zhang, Zijing Shi, Ling Chen, Mykola Pechenizkiy, and Jun Wang. 2024. Large language models are neurosymbolic reasoners. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17985–17993.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318.

Gaël Gendron, Jože M Rožanec, Michael Witbrock, and Gillian Dobbie. 2024. Counterfactual causal inference in natural language with large language models. *arXiv preprint arXiv:2410.06392*.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3023–3030.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and 1 others. 2023. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Dexter C Kozen. 1992. Depth-first and breadth-first search. In *The design and analysis of algorithms*, pages 19–24. Springer.

Yuangang Li, Yiqing Shen, Yi Nian, Jiechao Gao, Ziyi Wang, Chenxiao Yu, Shawn Li, Jie Wang, Xiyang Hu, and Yue Zhao. 2025. Mitigating hallucinations in large language models via causal reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16563–16577.

Haochen Liu, Song Wang, Yaochen Zhu, Yushun Dong, and Jundong Li. 2024. Knowledge graph-enhanced large language models via path selection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6311–6321.

Hang Luo, Jian Zhang, and Chujun Li. 2025. Causal graphs meet thoughts: Enhancing complex reasoning in graph-augmented llms. *arXiv preprint arXiv:2501.14892*.

Elan Markowitz, Krupa Galiya, Greg Ver Steeg, and Aram Galstyan. 2025. Kg-llm-bench: A scalable benchmark for evaluating llm reasoning on textualized knowledge graphs. *arXiv preprint arXiv:2504.07087*.

Moritz Miller, Bernhard Schölkopf, and Siyuan Guo. 2025. Counterfactual reasoning: an analysis of in-context emergence. In *Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032.

Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Yuan Sui, Shanshan Feng, Huaxiang Zhang, Jian Cao, Liang Hu, and Nengjun Zhu. 2022. Causality-aware enhanced model for multi-hop question answering over knowledge graphs. *Knowledge-Based Systems*, 250:108943.

Xingyu Tan, Xiaoyang Wang, Qing Liu, Xiwei Xu, Xin Yuan, and Wenjie Zhang. 2025. Paths-over-graph: Knowledge graph empowered large language model reasoning. In *Proceedings of the ACM on Web Conference 2025*, pages 3505–3522.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6076–6085.

Jiawei Wang, Da Cao, Shaofei Lu, Zhanchang Ma, Junbin Xiao, and Tat-Seng Chua. 2024. Causal-driven large language models with faithful reasoning for knowledge question answering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4331–4340.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Zeyu Wang. 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*, pages 143–151.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. 2024. Causality for large language models. *arXiv preprint arXiv:2410.15319*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. 2024. Causalbench: A comprehensive benchmark for causal learning capability of llms. *arXiv preprint arXiv:2404.06349*.

Appendix

A Algorithm Pseudocode

Algorithm 1 presents the pseudocode of the proposed framework.

Algorithm 1 Target-Aware Causal Graph Reasoning

Input : Question q
Output : Predicted answer $\hat{y}$
// 1. Setup (Sec. 3.1)
$(X, Y_s, Y_b, I) \leftarrow \Phi_{\text{extract}}(q)$ $\mathcal{F} \leftarrow \{X\}$; $\mathcal{V}_{vis} \leftarrow \{X\}$; $\mathcal{N}_{close} \leftarrow \emptyset$
// 2. Graph Expansion (Sec. 3.2)
while $Y_b \notin \mathcal{F}$ ***and*** *depth* $< D$ **do**
  // Generate candidates avoiding history
  $\mathcal{C} \leftarrow \Phi_{\text{expand}}(\mathcal{F} \mid \text{Avoid} = \mathcal{V}_{vis})$
  // Identify Close Hits via FGVA
  foreach $v \in \mathcal{C}$ **do**
    if $\Phi_{align}(v, Y_b) = \textsc{CloseHit}$ **then**
      $\mathcal{N}_{close} \leftarrow \mathcal{N}_{close} \cup \{v\}$
  // Prune to Top-K
  $\mathcal{F} \leftarrow \Phi_{\text{rank}}(\mathcal{C}, Y_b, K)$ $\mathcal{V}_{vis} \leftarrow \mathcal{V}_{vis} \cup \mathcal{F}$ Update Graph $\mathcal{G}$
// 3. Bridging (Sec. 3.2)
foreach $u \in \mathcal{N}_{close}$ **do**
  if $\Psi_{bridge}(u, Y_b)$ *is valid* **then**
    Add edge $(u \to Y_b)$ to $\mathcal{G}$
// 4. Extraction & Auditing (Sec. 3.3)
$\mathcal{P}^* \leftarrow$ Top-kPaths$(\mathcal{G}, X \to Y_b)$ **foreach** $p \in \mathcal{P}^*$ **do**
  // Check Premise & Counterfactuals
  if $V_{prem}(p) = 0$ ***or*** $V_{cf}(p) < \tau$ **then**
    Discard path p
// 5. Decision (Sec. 3.4)
$E_{\text{graph}} \leftarrow$ AggregatePaths$(\mathcal{P}^*)$ $E_{\text{LLM}} \leftarrow \Phi_{\text{reason}}(q, \mathcal{P}^*)$ $\hat{y} \leftarrow$ ResolveConflict$(E_{\text{graph}}, E_{\text{LLM}})$
return $\hat{y}$ mapped to options via I

B Dataset Construction Details

B.1 Overview

Most existing causal question answering resources are *context-dependent*, where each question is paired with a passage and models are evaluated on evidence extraction (Jin et al., 2023; Bondarenko et al., 2022). In contrast, we target *context-free causal direction prediction* with *graph-grounded explanations*: each instance is defined over explicit variables/events (X, Y) and the system predicts whether intervening on X makes Y *more* or *less* likely.

We focus on the directional subset with labels {more, less}. Some datasets include no effect as a distractor option, but we exclude NoEff instances as gold labels to keep the evaluation protocol consistent. Table 5 summarizes dataset statistics reports how often questions include *surface modifiers* (e.g., "more/less probability of Y").

Statistic	DDXPlus	WIQA	CauseNet
Domain	Medical	Science	Open-domain
N	200	212	100
More	100	99	50
Less	100	113	50
Avg. words	23.7	15.8	9.7
Neg. (%)	47.5	10.4	0.0

Table 5: **Dataset statistics.** Directional subset only (more/less). Negation via keyword match (e.g., *not/no/without*) on the question stem.

B.2 DDXPlus-CausalEffect: Directional Effect Prediction from Medical Records

We use the official DDXPlus release (Fansi Tchango et al., 2022), including condition metadata, evidence metadata, and patient-level predefined splits. Each instance specifies an intervention variable X (a patient evidence) and an outcome variable Y (a pathology), and asks how changing X affects the probability of Y.

Association-based labeling (Stats Mode). Our primary construction assigns labels using empirical association statistics computed from patient records. For each pathology Y and evidence X, we compute n_Y, n_X, n_{XY}, and N (total patients), and estimate:

$$p(Y \mid X) = \frac{n_{XY}}{n_X}, \qquad p(Y \mid \neg X) = \frac{n_Y - n_{XY}}{N - n_X}, \tag{12}$$

then define $\Delta = p(Y \mid X) - p(Y \mid \neg X)$. Using a margin threshold τ, we assign:

$$\texttt{label}(X, Y) = \begin{cases} \texttt{more}, & \Delta > \tau \\ \texttt{less}, & \Delta < -\tau \\ \texttt{discard}, & |\Delta| \le \tau. \end{cases} \tag{13}$$

These labels are treated as **data-driven directional proxies** (association-based) rather than fully identified causal effects; discarding near-zero cases avoids ambiguous instances when statistics are insufficient.

Question rendering. We render each pair (X, Y) into a WIQA-style question with answer choices

restricted to {more, less} (optionally keeping no effect only as a distractor). An optional LLM rewriting step can improve fluency while preserving explicit X and Y.

B.3 WIQA Directional Subset

WIQA (Tandon et al., 2019) is the closest widely-used benchmark aligned with our evaluation format because it is framed as effect-direction prediction and its questions are formulated over explicit standalone events. Following prior work, we evaluate on a curated subset that contains only directional-effect instances (more/less) and excludes NoEff cases.

B.4 CauseNet-derived Context-free Directional QA

We construct a general-domain dataset from CauseNet (Heindorf et al., 2020). Each record provides a head–tail causal assertion with optional confidence-like fields. We (i) sanitize concepts, (ii) normalize heterogeneous confidence signals to a unified edge confidence $c(e) \in [0, 1]$, and (iii) map relations to signed directed edges with $\text{sign} \in \{+1, -1\}$ (defaulting to $+1$ if polarity is missing).

Multi-hop path sampling and labeling. We sample simple directed paths $\pi = (x_0 \to \cdots \to x_h)$ and derive the gold direction label by polarity product:

$$\texttt{label}(\pi) = \begin{cases} \texttt{more}, & \prod_{e \in \pi} \text{sign}(e) = +1, \\ \texttt{less}, & \prod_{e \in \pi} \text{sign}(e) = -1. \end{cases} \tag{14}$$

We keep only directional instances to align with our evaluation protocol.

Question rendering. For each sampled path, we form a WIQA-style context-free question over endpoints $X = x_0$ and $Y = x_h$ and store the underlying sampled path as an auditable explanation trace.

Limitations. This construction inherits noise from open IE extraction and may underrepresent negative relations when explicit inhibitory edges are sparse; we therefore report label distributions and key sampling hyperparameters alongside results.

C Prompt Templates and Output Constraints

C.1 Common I/O Format

All datasets are cast into a unified **context-free directional-effect** format: each instance specifies explicit variables/events (X, Y) and asks whether intervening on X makes Y **more** or **less** likely. Following prior WIQA protocols, we exclude NoEff and keep a **two-choice** setting. Accordingly, all prompts in this paper share the same answer space: **A: more**, **B: less**. All LLM calls use deterministic decoding with temperature=0.

C.2 Answer Extraction and Fallback

For all methods, we enforce an explicit final answer line: Final answer: A or Final answer: B. We extract the choice using a regex matcher. If the required line is missing, we apply a lightweight *forced extractor* prompt that maps the model output to $\{A, B\}$; if extraction still fails, we fall back to A.

C.3 Baseline Prompts

Direct LLM. The model receives only the question stem and the two options, and must output the final choice:

```
[Direct] {question_stem}
A) more   B) less
Output format: Final answer: <A|B>
```

Chain-of-Thought (CoT). We elicit a brief rationale before producing the final choice, while keeping the same answer constraint:

```
[CoT] Write brief causal reasoning, then
choose A/B.
Output format:
Reasoning: <1-4 sentences>
Final answer: <A|B>
```

Graph-of-Thought (GoT). GoT follows a split $\to$ analyze $\to$ merge procedure with explicit intermediate artifacts. We first request a JSON decomposition into three components; then we analyze each component and produce a merged solution. All stages are deterministic and the final response must end with the forced choice:

```
Split:  Output  ONLY  JSON  {Component
1,2,3}.
Analyze each component.
Merge and MUST end with: Final answer:
<A|B>.
```

Tree-of-Thought (ToT). ToT performs a deterministic tree-style deliberation with explicit candidate generation and selection: (i) generate two candidate reasoning structures; (ii) select the better one; (iii) generate two candidate reasoning traces conditioned on the selected structure; (iv) select the better trace; (v) output the final choice. The final output is forced into $\{A, B\}$:

```
Generate 2 candidates (JSON) → Pick 1/2
Generate 2 traces (JSON) → Pick 1/2
Final: MUST end with: Final answer:
<A|B>
```

We refer to these as GoT/ToT because they enforce graph-/tree-structured intermediate representations and selection steps, even though decoding is deterministic in our implementation.

C.4 Our Framework: Graph-grounded + Audit Prompts

Our method uses the LLM as a *structured evidence worker* and a final *evidence aggregator*. All intermediate prompts are constrained to JSON outputs to ensure robust parsing, while the final decision is mapped to the unified two-choice space (A/B).

P1: Variable extraction (strict JSON).

```
You are a STRICT string-matching text
extractor.
Your job is ONLY to extract spans from
the given question text.
Do NOT paraphrase. Copy spans exactly.

Question:
"{question}"

Output ONLY valid JSON (no extra text):
{
  "cause_event": "<copied text>",
  "outcome_text_raw": "<copied text>",
  "outcome_direction": "MORE" or "LESS"
  or "NONE",
  "is_negated": true or false
}
```

P2: Target-aware single-hop expansion (forward).

```
You are a causal edge finder.

Input:
- CAUSE_NODE (X): "{X}"
- TARGET_HINT (Y): "{target_hint}"
- FORBIDDEN LIST (Avoid revisits):
[{avoid_str}]

Task:
- Propose up to {max_relations}
SINGLE-HOP causal effects
starting from X.
- If TARGET_HINT is not "NONE",
expand toward Y (prefer intermediates
that connect X to Y).
- Each tail must be a NEUTRAL NOUN
PHRASE (no "more/less", no full
sentences).
- Use "INCREASES" when increasing
head tends to increase tail.
- Use "DECREASES" when increasing
head tends to decrease tail.

Output ONLY JSON:
{
  "triples": [
    ["{X}", "INCREASES" | "DECREASES",
    "<neutral noun phrase>"]
  ]
}
```

P3: Target equivalence / bridging judgment.

```
You are judging the relationship
between two variables in a causal system.

Variable A: "{A}"
Variable B: "{B}"

Task:
Decide their relationship along
three axes:
1) core_entity_relation
2) quantity_relation
3) causal_or_structural_relation

Output ONLY JSON:
{
  "core_entity_relation": "...",
  "quantity_relation": "...",
  "causal_or_structural_relation":
  "...",
  "explanation": "short explanation"
}
```

P4: Counterfactual edge audit (binary validity).

```
You are a Scientific Logic Judge.

We MUST judge causality based on
```

```
intervention semantics:
If we actively increase A,
does B tend to increase/decrease?

Candidate causal edge:
A: "{A}"
Relation: "{REL}"  (INCREASES/DECREASES)
B: "{B}"

Return a conservative judgment. If unsure,
mark false.

Output ONLY JSON:
{"is_valid_link": true/false, "reasoning":
"short explanation"}
```

P5: Final aggregation with audited chains.

```
You are solving a WIQA-style causal
reasoning problem.
Your job is to decide how the CAUSE affects
the BASE VARIABLE.

Question: "{question}"
Cause event (X): "{cause_event}"
Outcome event (surface): "{outcome_event}"
BASE variable (outcome_base,
the only quantity you judge):
"{outcome_base}"

Causal graph summary:
{summary_json}

Evidence chains from cause →
BASE (system-computed net effects;
DO NOT re-multiply signs yourself):
{evidence_block}

IMPORTANT:
- Decide the direction of change for
the BASE VARIABLE only ("{outcome_base}").
- [Net Effect: POSITIVE (Causes Increase)]
supports "more" (A).
- [Net Effect: NEGATIVE (Causes Decrease)]
supports "less" (B).
- If chains conflict, prefer
higher-quality / fewer-bridge chains.

Output ONLY strict JSON:
{
  "effect_on_base": "more" | "less",
  "final_answer": "A" | "B",
  "confidence": "high" | "medium"
  | "low" | "very_low",
  "reasoning": "short explanation grounded
  in the evidence chains"
}
```

Example evidence_block (serialized chains).

```
- Chain: X -> INCREASES -> M1 ->
DECREASES -> outcome_base
  [Net Effect: NEGATIVE (Causes Decrease)]
  [bridge_edges: 0/2]
- Chain: X -> INCREASES -> M2 -> INCREASES
-> outcome_base
  [Net Effect: POSITIVE (Causes Increase)]
  [bridge_edges: 1/2]
```

Method	Acc. (%)
Direct	53.69
Ours	**67.54**

Table 6: Results on a larger WIQA slice with 650 instances.

τ_{audit}	0.4	0.5	0.6	0.7	0.8
Path Reach (%)	97.51	97.23	96.23	86.18	84.48
Acc. (%)	62.11	65.82	**67.92**	65.43	63.03

Table 7: Sensitivity to the audit threshold τ_{audit} on WIQA.

D Additional Experimental Analyses

D.1 Robustness to Evaluation Size

To test whether the WIQA gain persists beyond the 212-example evaluation subset, we additionally evaluate on a larger slice of 650 instances under the same task and labeling protocol. The performance gap remains substantial, confirming that the gain is not an artifact of a small evaluation set.

D.2 Sensitivity to the Audit Threshold and Bridge Penalty

We further analyze sensitivity to the audit threshold τ_{audit} and the bridge penalty γ on WIQA. Increasing τ_{audit} enforces stricter filtering and lowers Path Reach; $\tau_{\text{audit}} = 0.6$ provides the best balance between accuracy and coverage. Moderate γ performs best, whereas small γ admits too many bridge edges and large γ over-penalizes bridging.

D.3 Efficiency and Parallelization

We report average per-instance wall-clock time and token usage on WIQA. The framework is naturally

γ	0.2	0.4	0.6	0.8
Acc. (%)	64.23	66.83	**67.92**	62.81

Table 8: Sensitivity to the bridge penalty γ on WIQA.

Method	Avg. Time (s)	Avg. Tokens
Direct	0.698	63.54
CoT	2.545	268.21
ToT	18.664	3928.56
Causal-Audit (ours)	136.166	42656.52
Causal-Audit (8 threads)	21.324	42685.31

Table 9: Efficiency and token usage on WIQA.

parallelizable because frontier expansion and counterfactual auditing can be executed independently across candidate nodes, edges, and paths. Eight-thread execution substantially reduces wall-clock latency while leaving token usage essentially unchanged.

D.4 Robustness of Counterfactual Auditing

We assess three robustness properties of counterfactual edge auditing: seed consistency, prompt paraphrase sensitivity, and reversed-direction checks. We randomly sample 50 questions and extract 339 audited edges from their constructed graphs.

Seed consistency. With temperature set to 0, varying the random seed does not affect $V_{cf}(e)$ or the final outputs in our implementation. Across five seeds, the variance of $V_{cf}(e)$ is 0 for all 339 audited edges.

To test directional robustness, we re-audit the reversed direction ($B \rightarrow A$) for the same set of 339 audited forward edges ($A \rightarrow B$) under the same protocol. Among these 339 edges, 77 reversed directions pass the audit, yielding a reverse-direction pass rate of 22.71% (77/339). This low reverse-direction pass rate indicates that the audit is direction-sensitive, as expected for causal relations.

D.5 Decoupled Generation and Auditing

To test whether auditing provides an error-correction signal independent of the generation backbone, we compare a coupled setting (gen=audit=Llama-3.1-8B) with a decoupled setting (gen=Llama-3.1-8B, audit=Mistral-7B). Accuracy decreases slightly from 67.92 to 65.13, suggesting that under the current configuration we do not observe a stable generation-independent correction benefit.

Range of σ_e for $V_{cf}(e)$	# Edges	% of all
$[0, 0.1]$	262	77.29
$(0.1, 0.2]$	42	12.39
$(0.2, 0.3]$	19	5.60
$(0.3, 0.4]$	11	3.24
$(0.4, 0.5]$	5	1.47

Table 10: Distribution of per-edge standard deviation σ_e of $V_{cf}(e)$ across prompt paraphrases.

D.6 Failure Analysis

We provide a representative failure case to localize the dominant error source in our pipeline.

Instance. Question: "Suppose if the helix increases and divides in parts of 3 happens, how will it affect hurting the DNA to replicate properly."
Intervention variable: $X =$ *helix increases and divides in parts of 3*.
Base target: $Y_b =$ *hurting the DNA to replicate properly*.
Gold label: `more`.
Prediction: `less`.

Audited causal paths. The model retains two high-scoring but semantically inconsistent paths:

$$\text{(P1)}\quad X \xrightarrow{\text{DEC}} \texttt{DNA replication fidelity} \xrightarrow{\text{DEC}} \texttt{DNA damage} \xrightarrow{\text{INC}} \texttt{genomic stability} \xrightarrow{\text{DEC}} Y_b,$$

$$\text{(P2)}\quad X \xrightarrow{\text{DEC}} \texttt{DNA replication fidelity} \xrightarrow{\text{DEC}} \texttt{DNA damage} \xrightarrow{\text{INC}} Y_b.$$

Error source. The dominant failure source is *semantic drift* at the intermediate node `DNA damage`. In P1, the model no longer interprets `DNA damage` literally as physical lesions or harm. Instead, it implicitly shifts toward a damage-response or repair-related state, which then supports the chain *more repair / more genomic stability / less replication failure*, producing an overall `less` effect on the target. In P2, the same phrase is interpreted literally as actual damage, in which case more damage directly supports more replication failure, which is consistent with the gold label `more`.

Implication. This error does not primarily arise from missing graph connectivity or the absence

of audited paths. Instead, it arises because the same intermediate node is used with incompatible senses across surviving paths, allowing the aggregated sign to flip. This suggests that, beyond edge-level auditing, future improvements should enforce stronger *path-internal semantic consistency* for reused intermediate variables.

E Extended Case Study

Instance. Intervention: X =*the patient has not noticed any new fatigue, vague discomfort, diffuse muscle aches, or a change in well-being*. Target: Y_b =*Chagas probability*. Gold label: `less`.

Baseline (CoT) output. CoT predicts `more` by invoking a generic narrative that Chagas can be asymptomatic in early stages, treating symptom absence as evidence for infection.

Representative generated causal triples. Our graph construction produces localized, inspectable hypotheses as directed triples. Examples include:

- $X \xrightarrow{\text{DEC}}$ `patient's overall health status`
- $X \xrightarrow{\text{INC}}$ `Chagas disease risk factors`
- `Chagas disease risk factors` $\xrightarrow{\text{INC}}$ `infection probability`
- `infection probability` $\xrightarrow{\text{DEC}}$ `Chagas probability`
- `Chagas disease risk factors` $\xrightarrow{\text{INC}}$ `Chagas probability`
- `patient's overall health status` $\xrightarrow{\text{INC}}$ `Chagas probability`

Extracted causal chains. From the constructed graph, our path extractor returns three chains:

(P1) $X \xrightarrow{\text{DEC}}$ `overall health status` $\xrightarrow{\text{INC}} Y_b$,
(P2) $X \xrightarrow{\text{INC}}$ `risk factors` $\xrightarrow{\text{INC}} Y_b$,
(P3) $X \xrightarrow{\text{INC}}$ `risk factors` $\xrightarrow{\text{INC}}$ `infection probability` $\xrightarrow{\text{DEC}} Y_b$.

Counterfactual edge audit. For each edge $e = (u, v)$ on a candidate path, we compute a counterfactual support score $V_{\text{cf}}(e) \in [0, 1]$ via intervention-style probing, and define the path audit score as

$$S_{\text{audit}}(p) = V_{\text{prem}}(p) \cdot \prod_{e \in p} V_{\text{cf}}(e). \qquad (15)$$

In this example, the retained edges receive consistently high counterfactual support (approximately 0.9–1.0), and the final decision is made by aggregating the audited paths.

Aggregation and decision. The audited chains contain directional disagreement (one positive vs. two negative). Our conflict-aware aggregation weights each chain by its audit score and combines positive/negative evidence mass, resulting in a net `less` effect on Y_b and a final confidence of 0.85, matching the gold label.

F Ethics Statement

This manuscript is the authors' original work. Except for minor English grammar checking with ChatGPT, no large language model or AI tool was used for idea generation, problem formulation, literature search or screening, methodology design, code implementation, data processing, experimental design, statistical analysis, figure or table drafting, or substantive writing. All intellectual contributions, including conceptualization, model design, and empirical evaluation, are solely those of the authors.