

METER: Evaluating Multi-Level Contextual Causal Reasoning in Large Language Models

Pengfeng Li^{1,3} Chen Huang^{2*} Chaoqun Hao¹ Hongyao Chen¹ Xiao-Yong Wei¹
Wenqiang Lei^{1,3} See-Kiong Ng²

¹ College of Computer Science, Sichuan University

² Institute of Data Science, National University of Singapore

³ Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, China
lipengfeng109@gmail.com, huang_chen@nus.edu.sg

Abstract

Contextual causal reasoning is a critical yet challenging capability for Large Language Models (LLMs). Existing benchmarks, however, often evaluate this skill in fragmented settings, failing to ensure context consistency or cover the full causal hierarchy. To address this, we pioneer METER to systematically benchmark LLMs across all three levels of the causal ladder under a unified context setting. Our extensive evaluation of various LLMs reveals a significant decline in proficiency as tasks ascend the causal hierarchy. To diagnose this degradation, we conduct a deep mechanistic analysis via both error pattern identification and internal information flow tracing. Our analysis reveals two primary failure modes: (1) LLMs are susceptible to distraction by causally irrelevant but factually correct information at lower level of causality; and (2) as tasks ascend the causal hierarchy, faithfulness to the provided context degrades, leading to a reduced performance. We believe our work advances our understanding of the mechanisms behind LLM contextual causal reasoning and establishes a critical foundation for future research. Our code and dataset are available at <https://github.com/SCUNLP/METER>.

1 Introduction

Deriving causal conclusions by grounding reasoning in a specific natural language context, a process we term **Contextual Causal Reasoning**, is considered a fundamental prerequisite for achieving Artificial General Intelligence (Feder et al., 2022; Yang et al., 2022; Wang, 2024; Joshi et al., 2024; Saklad et al., 2025). This is particularly critical in high-stakes domains such as medical diagnosis, which typically involve diverse causal inquiries (Richens et al., 2020; Wu et al., 2024). For instance, a clinical report analysis (Figure 1) requires

diverse causal inquiries: discovering the causes of ischemia, predicting the outcomes of interventions (e.g., *What if PCI is performed?*), and simulating counterfactual to review diagnostic errors.

With Large Language Models (LLMs) demonstrating superior efficacy on numerous tasks, a growing body of research has emerged to benchmark their proficiency in contextual causal reasoning (Bondarenko et al., 2022; Ho et al., 2023; Romanou et al., 2023). However, existing benchmarks remain fragmented investigation (Yang et al., 2024). Specifically, 1) **Incomplete Coverage of Assessment Levels**. Causal reasoning spans three essential levels (Pearl and Mackenzie, 2018): causal discovery (i.e., association), intervention, and counterfactual. Yet existing benchmarks focus exclusively on either causal discovery (Tan et al., 2023; Su et al., 2025) or counterfactual reasoning (Yu et al., 2023). This narrow focus neglects the necessity for cross-perspective consistency. Consequently, they fail to reveal performance disparities across the different levels of causal reasoning. 2) **Biased Evaluation due to Contextual Inconsistency**. Constructing a fair multi-level assessment presents significant challenges, as it necessitates generating questions across all three causal layers derived from a single, unified context to ensure a controlled comparison. Yet, existing benchmarks, including those not explicitly designed for contextual causal reasoning, typically employ disparate contexts for different questions, preventing a reliable evaluation (Jin et al., 2023; Wang, 2024).

To this end, we introduce **METER**, a benchmark tailored for the Multi-Level assessment of contextual causal reasoning in LLMs. METER incorporates two primary contributions. First, inspired by the Ladder of Causation (Pearl and Mackenzie, 2018), METER is structured around the three levels: causal discovery, intervention, and counterfactual. Each level is properly defined for causal analysis in the textual domain (cf. Section

*Corresponding author.

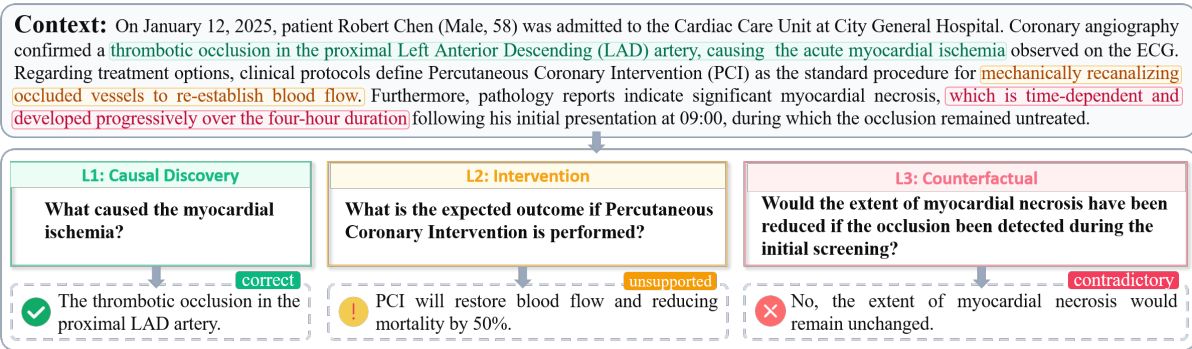


Figure 1: Illustration of Contextual Causal Reasoning.

3.1), forming a structured framework for evaluating context-based causal reasoning. Second, leveraging this structure, we curate a benchmark dataset of 4,145 entries that enforces contextual consistency: every entry features a unique context paired with questions spanning all three causal levels. This rigor mitigates confounding factors from context variation, facilitating a reliable evaluation of LLMs’ across-level performance.

Our findings reveal a distinct performance hierarchy: despite demonstrating strong proficiency in causal discovery, LLM performance degrades significantly as the complexity of the causal level increases. Specifically, we observe substantial drops in intervention (-15.78%) and counterfactual reasoning (-26.27%). Furthermore, our error analysis indicates that failures in Discovery are predominantly driven by distractions from irrelevant background details, whereas higher-level tasks suffer from a lack of faithfulness to the provided context. To gain a deeper understanding of these tendencies, we conduct a mechanistic investigation, analyzing the dynamics of information flow across LLM layers through the saliency technique (Wang et al., 2023). To sum up, we believe that METER stands as a comprehensive resource to rigorously benchmark and decipher the causal reasoning mechanisms of LLMs. To sum up, our main contributions are as follows:

- We propose METER, a benchmark for the multi-level assessment of contextual causal reasoning in LLMs, which resolves the limitations of incomplete coverage and evaluation bias.
- We contribute a dataset comprising 4,145 entries that enforces contextual consistency, providing a reliable evaluation foundation for future research.
- We evaluate the strengths and weaknesses of var-

ious LLMs, revealing their proficiency and error patterns across varying causal levels.

- We pioneer the mechanistic investigation of LLMs, revealing the internal information flow dynamics that drive their reasoning behaviors.

2 Related Work

Contextual Causal Reasoning. Unlike common-sense causal reasoning, which only leverages the intrinsic world knowledge of LLMs to infer causal conclusions (Miliiani et al., 2025; Frohberg and Binder, 2022; Lasheras and Pinheiro, 2025), or formal causal reasoning, which is grounded in symbolic logic and well-defined formal rules (Jin et al., 2023; Wang, 2024), contextual causal reasoning is evidence-driven¹ (cf. Table 1). It necessitates the extraction and synthesis of causal chains directly from the provided narrative, requiring the model to align its reasoning strictly with the situational evidence rather than general priors (Dalal et al., 2023; Kiciman et al., 2023). Given the critical role of the contextual causal reasoning in building robust and human-like AI (Lake et al., 2017; Chen et al., 2024; Ma, 2025), numerous studies have sought to assess how well LLMs perform this task. However, existing benchmarks restrict their scope to a single aspect of the causal levels (Chi et al., 2024; Cheng et al., 2025a). For instance, some focus on causal discovery (Ho et al., 2023; Bondarenko et al., 2022; Tan et al., 2023; Romanou et al., 2023), while others are limited to counterfactual reasoning (Yu et al., 2023). This fragmentation precludes evaluating LLMs across distinct causal levels within a unified narrative, hindering a comprehensive understanding of their contextual

¹cf. illustration example in Appendix A for difference.

Benchmark	Reasoning Paradigm	Causal Discovery	Intervention	Counterfactual	Unified Context	Mechanistic Analysis
ExpliCa (Miliari et al., 2025)	Commonsense	✓	✗	✗	✗	✗
CRASS (Frohberg and Binder, 2022)	Commonsense	✗	✗	✓	✗	✗
CalQuest (Lasheras and Pinheiro, 2025)	Commonsense	✓	✓	✓	✗	✗
CLADDER (Jin et al., 2023)	Formal	✓	✓	✓	✗	✗
CausalBench (Wang, 2024)	Formal	✓	✓	✓	✗	✗
WIKIWHY (Ho et al., 2023)	Contextual	✓	✗	✗	✗	✗
CausalQA (Bondarenko et al., 2022)	Contextual	✓	✗	✗	✗	✗
RECESS (Tan et al., 2023)	Contextual	✓	✗	✗	✗	✗
CRAB (Romanou et al., 2023)	Contextual	✓	✗	✗	✗	✗
IfQA (Yu et al., 2023)	Contextual	✗	✗	✓	✗	✗
METER (Ours)	Contextual	✓	✓	✓	✓	✓

Table 1: Comparison of various benchmarks. *Unified Context* highlights the contextual inconsistency of current benchmarks. *Mechanistic Analysis* reveals their failure to understand internal reasoning mechanisms of LLMs.

causal reasoning capabilities. We propose a benchmark for the multi-level assessment of contextual causal reasoning in LLMs. It features a unified context, paired with questions across all causal levels, reducing contextual confounds and enabling fair cross-level evaluation.

Mechanistic Interpretability. Mechanistic interpretability aims to uncover the internal mechanisms that drive the behaviors of LLMs by inspecting individual components and their interactions (Zhao et al., 2024; Luo and Specia, 2024). Prominent approaches in this domain include vocabulary lens (Geva et al., 2022; Belrose et al., 2023; Ortu et al., 2024), causal tracing (Meng et al., 2022), and circuit discovery (Wang et al., 2022a, 2023; Ferrando and Voita, 2024). Notably, the saliency technique has been effectively employed to capture the interaction dynamics between tokens, revealing how input tokens contribute to the model’s prediction (Wang et al., 2023). In this work, we adopt this method to investigate the LLMs’ working mechanism driving contextual causal reasoning.

3 METER Benchmark

METER operates on a multi-level causal framework spanning causal discovery, intervention, and counterfactual reasoning. Based on this hierarchy, we curate a dataset of 4,145 samples, where each entry features a unified context paired with questions targeting all three causal levels. To ensure robust evaluation, questions are designed as multiple-choice items with five options. We present the dataset statistics in Table 2 and detail on the construction pipeline below.

	Level 1	Level 2	Level 3
# of Contexts	4,145	4,145	4,145
# of Questions	4,145	4,145	4,145
# of Options per Question	5	5	5
Avg. # Tokens per Context	228.91	228.91	228.91
Avg. # Tokens per Question	18.77	26.78	26.16
Avg. # Tokens per Option	17.62	20.28	20.69

Table 2: Data statistics across different causal levels.

3.1 Causal Reasoning Levels

Pearl’s Ladder of Causation (Pearl and Mackenzie, 2018) provides the theoretical foundation for multi-level causal reasoning. However, it is formulated for numerical and probabilistic domains. Natural language reasoning, by contrast, necessitates a reliance on semantic inference. Inspired by Yang et al. (2024), we transpose these causal levels into the linguistic paradigm, defining them as follows.

- **Level 1: Causal Discovery.** This level entails identifying causal relations expressed within the text, disentangling causation from mere association (Cheng et al., 2025b). LLMs must detect relations signaled by explicit lexical cues (e.g., ‘caused’, ‘because’) or implicit semantic context. As illustrated in Figure 1, LLMs must leverage the explicit verb ‘causing’ to map the cause (‘thrombotic occlusion’) to the effect (‘myocardial ischemia’).
- **Level 2: Intervention.** This level involves forecasting the consequences of introducing a new action within a given context. LLMs must leverage the text’s causal logic to infer how a novel event alters the outcome. As illustrated in Figure 1, the intervention question regarding *Percutaneous Coronary Intervention*

(PCI) demands multi-step reasoning: *performe PCI*→*clear occlusion*→*blood flow restores*. This task tests a deeper capability than mere causal discovery, requiring the LLM to dynamically apply causal rules established within the context to predict the impact of external changes.

- **Level 3: Counterfactual.** This level entails reasoning about alternative outcomes by modifying past events. LLMs must infer causal outcomes under hypothetical conditions that differ from the factual context. As illustrated in Figure 1, the counterfactual question (e.g., L3) demands that the LLM reverse the known condition and infer that damage would be prevented. This differs from Level 2, which predicts future outcomes based on the existing reality, whereas Level 3 requires the ability to simulate ‘what-if’ scenarios that conflict with the provided contextual facts.

3.2 Benchmark Data Construction

We adopt a human-LLM collaboration approach to construct the dataset. Finally, each data entry features a unified context paired with questions targeting all three causal levels. To ensure robust evaluation, questions are designed as multiple-choice items with five options. More details on data construction are provided in Appendix B.

Data Preparation. We source our initial data from four established datasets: ESL (Caselli and Vossen, 2017), MAVEN-ERE (Wang et al., 2022b), MECI (Lai et al., 2022), and WIKIWHY (Ho et al., 2023). Each instance comprises a passage context accompanied by annotated cause-effect event pairs. Notably the first three datasets describe events using only brief trigger words. These triggers often lack the semantic completeness required to construct natural and unambiguous causal questions. To address this, we employ Gemini 2.5-Pro to expand these triggers into full event descriptions, where a three-stage quality control process is utilized to mitigate potential hallucinations or ambiguity. This filters out over 90% of the initial pool, yielding 6372 high-quality, context-dependent cause-effect pairs for subsequent data generation.

- **Length-based Filtering.** An automated filter discards descriptions shorter than three tokens.
- **Data De-contamination.** To avoid cases where LLMs have memorized the cause-effect relationships, we adopt a filtering strategy inspired by Xiao et al. (2025). We prompt an ensemble of

advanced LLMs² to evaluate the causal relationship between pairs *in the absence of context*. We retain only those pairs where all models unanimously agree that no inherent causal link exists.

- **Human Verification.** Three human annotators manually review the remaining entries to remove ambiguous, erroneous or semantically incomplete event descriptions, discarding any instance rejected by at least one annotator. We measure the inter-annotator agreement using Fleiss’s κ (Landis and Koch, 1977), achieving a score of 0.78, which reflects a high level of consistency.

Data Generation. Inspired by Guo and Chen (2024); Long et al. (2024), we adopt Gemini 2.5-Pro for generation. For each cause-effect pair, we generate three distinct multiple-choice causal questions (causal discovery, intervention, counterfactual), along with their respective answer choices.

- **Question Generation.** We design multiple templates tailored for different levels to generate causal questions³. Specifically, for causal discovery, we instantiate templates directly with the cause or effect event (e.g., *Why does EVENT happen*). For intervention and counterfactual, we prompt Gemini 2.5-Pro to propose diverse hypothetical conditions based on the given cause-effect pair, which are then applied to the corresponding templates. Finally, all filled templates are paraphrased by Gemini 2.5-Pro to ensure natural and unambiguous phrasing.
- **Answer Generation.** For each question, we generate the corresponding correct answer. For causal discovery questions, the answer is taken directly from the annotated cause or effect event. For intervention and counterfactual questions, we prompt Gemini 2.5-Pro to generate the expected outcome, conditioned on the ground-truth cause-effect pairs and specific reasoning guidelines.
- **Distractor Generation.** We design various distractors to systematically investigate different weaknesses in causal reasoning. Specifically, we define four categories of distractors, as shown in the Table 3. Given a question and its correct answer, Gemini 2.5-Pro is prompted to generate candidates from these categories under strict

²Gemini 2.5-Pro, GPT-5, Qwen3-235b-a22b-instruct-2507

³Detailed in Table 9 in Appendix B.2.

Category	Definition
Contradictory Statement	Options conflicting with facts in the context or premises in the question.
Unfounded Statement	Options containing information that is not stated in or cannot be inferred from the context.
Causal Reversal	Options inverting the cause-and-effect direction established in the context.
Irrelevant Fact	Options that are grounded in the context with no causal link to the question.

Table 3: Definitions of the four distractor categories.

format constraints. This controlled generation ensures that distractors remain semantically coherent with the passage while introducing specific pitfalls that can mislead models.

Human Validation & Quality Control. To guarantee dataset reliability, we employ a human validation process, editing and filtering any instances that deviate from our established quality benchmarks. This yields a final curated dataset of 4145 high-quality samples.

- **Manual Editing.** We ask a group of NLP-background annotators to edit and refine the generated samples. The revision targets: (1) *Questions*. Rewriting unclear or unnaturally phrased questions to improve fluency. (2) *Correct Options*. Verifying that the correct answer is both accurate and comprehensive, refining the text for clarity. (3) *Distractors*. Ensuring each distractor conforms to its assigned category and revising any that are ambiguous or misaligned.
- **Manual Filtering.** To further enhance quality, another group of annotators is required to eliminate low-quality edited samples. Any sample failing to meet the following standards is discarded: (1) *Question Alignment*. The question must strictly correspond to its designated level of causal reasoning. (2) *Faithfulness*. Both the question and the correct answer must be fully grounded in the provided context, avoiding any unsupported external information. (3) *Fluency*. The text of the question and options must be grammatically correct, natural, and unambiguous. (4) *Accuracy*. The designated correct option must provide an accurate and comprehensive answer derived solely from the context. (5) *Distractors Alignment*. Each distractor must logically align with its predefined error category.

To ensure the reliability and consistency of our annotations, each sample is independently reviewed

by three annotators during both the filtering and editing stages, following Bender and Friedman (2018). The final retained samples are determined by a majority vote among the three annotations. We measure inter-annotator agreement using Fleiss’s κ and achieved a score of 0.71 for manual editing and 0.75 for manual filtering, indicating substantial agreement. Dataset samples and more details are listed in Appendix E.1 and B.3, respectively.

4 Experiment

We guide our evaluation with three research questions: **RQ1**: How do LLMs perform across the three levels of causal reasoning? **RQ2**: How do LLMs’ error patterns vary across different causal levels? **RQ3**: How can the observed error characteristics be explained? Refer to Appendix D and E.2 for more experiments and case studies.

4.1 Experiment Setup

Evaluated LLMs & Prompting Schemes. We benchmark various LLMs encompassing both closed-source and open-source models with varying specializations. For the closed-source LLMs, we evaluate reasoning-optimized models (*GPT-5* and *Gemini3-Pro*) alongside instruction-tuned models (*GPT-4o* and *Gemini3-Flash*). For the open-source LLMs, we select leading instruction-tuned models (*Qwen3-Next-80B-A3B-Instruct* and *Llama-3.3-70B-Instruct*) and the reasoning-optimized model (*Qwen3-Next-80B-A3B-Thinking*). To investigate scaling laws, we include the *Qwen3* family of different sizes. Additionally, we use four prompting paradigms: *Zero-shot*, *Few-shot*, *Zero-shot CoT*, and *Few-shot CoT*. Specifically, instruction-tuned models are tested under all four conditions, whereas reasoning-optimized models are evaluated in *Zero-shot* and *Few-shot* modes, given their intrinsic reasoning capabilities. See Table 21 for detailed specifications.

Evaluation Metrics. Following (Chi et al., 2024), we adopt Accuracy as the metric for evaluation. Refer to Appendix C.2 for implementations.

4.2 Overall Performance (RQ1)

We report the performance of various LLMs across the three causal levels in Table 4. All results reported in the table are averaged over three independent runs. Our detailed observations are as follows. **Model proficiency significantly diminishes across ascending causal levels.** Table 4 reveals

Model	Causal Discovery				Intervention				Counterfactual			
	ZS	CoT	FS	FS+C	ZS	CoT	FS	FS+C	ZS	CoT	FS	FS+C
GPT-4o	87.92	86.46	87.01	87.94	77.96	75.14	78.61	79.45	67.42	66.18	68.99	68.67
Gemini3-Flash	91.02	89.52	91.29	90.55	78.09	78.83	81.17	81.21	70.51	71.08	73.78	72.44
Qwen3-Next (Instruct)	89.56	85.83	85.98	88.46	75.13	76.40	74.20	78.68	64.47	66.29	66.32	68.62
Llama-3.3-70B-Instruct	87.24	85.15	88.40	86.76	78.17	79.33	80.87	77.79	62.08	67.36	70.78	69.55
GPT-5	92.96	–	92.93	–	82.17	–	83.80	–	72.14	–	73.99	–
Gemini3-Pro	93.50	–	93.28	–	81.92	–	84.69	–	73.05	–	77.49	–
Qwen3-Next (Thinking)	90.36	–	90.46	–	77.52	–	82.52	–	70.57	–	73.51	–
Qwen3-0.6B	64.46	59.41	51.81	55.18	31.27	24.84	28.64	25.69	25.88	20.99	23.78	21.23
Qwen3-4B	87.03	85.54	85.01	85.61	53.47	53.12	59.81	56.12	43.26	42.48	44.24	39.19
Qwen3-8B	86.27	82.64	86.40	86.42	64.48	64.34	66.25	67.15	51.40	53.08	51.32	53.27
Qwen3-14B	87.95	86.78	87.42	87.00	67.54	69.90	70.67	71.45	52.05	56.78	53.93	55.82
Qwen3-32B	87.26	86.53	87.05	86.58	71.54	74.39	73.46	74.50	61.12	62.79	62.67	61.07
Human		95.80				92.80				91.00		

Table 4: Overall performance across three causal levels with different prompting strategies. **ZS**: Zero-shot, **CoT**: Zero-shot CoT, **FS**: Few-shot, **FS+C**: Few-shot CoT. **Bold** denotes the best result across all models. Missing values (–) indicate strategies not applicable to reasoning-optimized models.

that while LLMs are highly effective at causal discovery (e.g., Gemini3-Pro reaches 93.50% accuracy), their proficiency falters on more complex tasks. Specifically, Gemini3-Pro suffers significant drops in intervention (81.92%) and counterfactual (73.05%) scenarios. This discrepancy underscores that while current LLMs excel at extracting explicit causal patterns from text, they struggle with the advanced reasoning required to simulate outcomes under intervention or hypothetical negation.

Reasoning-optimized models exhibit superior robustness in higher causal level. Comparative analysis reveals that while instruction-tuned models (e.g., Qwen3-Next-Instruct) are competitive with reasoning-optimized models (e.g., Qwen3-Next-Thinking, Gemini3-Pro) on causal discovery, they lack robustness on advanced tasks. The performance divergence in the few-shot setting is stark: Qwen3-Next-Thinking achieves 72.51% accuracy on counterfactual, markedly outstripping Qwen3-Next-Instruct (66.32%) and GPT-4o (68.99%). These results confirm that reasoning optimization confers a significant capability for dynamic causal inference, attenuating the steep performance decay typically observed as tasks move from association to counterfactual simulation.

The efficacy of prompting strategies varies. According to Table 4, prompting efficacy can be non-uniform, heavily influenced by both the causal hierarchy and model architecture. For higher-level tasks (intervention, counterfactual), few-shot prompting emerges as a robust strategy,

consistently enhancing performance (e.g., +3.6% for Gemini3-Pro). The role of CoT, however, is architectural-dependent. It provides critical scaffolding for Llama-3.3-70B-Instruct (62.08% → 67.36%), yet proves detrimental to GPT-4o (77.96% → 75.14%). We hypothesize that while Few-shot demonstrations provide necessary contextual grounding, explicit CoT steps can introduce reasoning noise in models that have already internalized efficient solution paths.

Size Effect. As shown in Table 4, LLMs’ contextual causal reasoning capabilities generally scale with parameter count, yet the impact of model size varies distinctively across different causal levels. For causal discovery, performance stabilizes early; the 4B model (87.03%) achieves accuracy comparable to the 32B model (87.26%), indicating that identifying causal relations expressed within the text is achievable even with limited parameters. In contrast, performance on Intervention and Counterfactual tasks exhibits a continuous improvement as model size increases. From 0.6B to 32B, we observe substantial gains (e.g., +49.54% for intervention and +41.80% for counterfactual). This suggests that the capabilities needed for higher-level tasks rely heavily on larger model scale.

Human performance. Human evaluators demonstrated robust proficiency across the causal levels⁴, achieving average accuracies of 95.8% on causal discovery, 92.8% on intervention, and 91.0% on counterfactual reasoning. These high scores con-

⁴cf. Appendix C.1 for details.

Error Type	Gemini3-Flash (%)			Qwen3-4B (%)		
	Discovery (N=424)	Intervention (N=953)	Counterfactual (N=1395)	Discovery (N=536)	Intervention (N=2046)	Counterfactual (N=2489)
Irrelevant Fact	45.56	21.40	16.49	55.41	26.87	24.15
Unfounded Statement	28.89	45.94	50.54	22.20	39.43	36.77
Contradictory Statement	12.00	18.05	26.28	5.69	20.96	33.87
Causal Reversal	13.56	14.60	6.69	16.70	12.74	5.22

Table 5: Error analysis (error percentage) across different causal levels under the zero-shot setting. The total number of incorrect samples N is provided in parentheses. Refer to Appendix C.3 for more results.

firm that METER aligns closely with human causal reasoning standards. Furthermore, it underscores a performance gap between current LLMs and human capabilities in these tasks.

4.3 Error Analysis (RQ2)

This section conducts an in-depth analysis to characterize the distinct error patterns across different levels of the context-based causal reasoning.

Setup. Our analysis centers on Gemini3-Flash and Qwen3-4B under the zero-shot protocol, as this captures the models’ intrinsic reasoning tendencies (Miao et al., 2024). To characterize failure modes, we manually categorize every erroneous prediction according to the predefined distractors in Table 3. We then calculate the proportion of each error type to identify dominant failure patterns. Further analyses of other LLMs and prompting schemes are documented in Appendix C.3. Finally, Table 5 summarizes the results.

Errors in causal discovery stem primarily from distraction by irrelevant contextual noise. As shown in Table 5, causal discovery errors are heavily concentrated in the *Irrelevant Fact* category (55.41% for Qwen3-4B; 45.56% for Gemini3-Flash). For example, when queried about the cause of a *car crash*, LLMs bypass the actual cause (*‘brake failure’*) in favor of explicit but non-causal narrative details (*‘the car was driving on Main Street’*). While the selected answers are true statements within the context, they serve as mere irrelevant contextual noises rather than actual causes.

Errors in higher-level tasks stem primarily from a diminishing faithfulness to the context, manifesting as unfounded and contradictory statements. Analysis of error distributions reveals a shift to faithfulness failures at higher causal levels (intervention and counterfactual). Errors are characterized by a detachment from the source text and a reliance on hallucinated information, evidenced by the dominance of Unfounded Statements (peak-

ing at 50.54% for Gemini3-Flash). Furthermore, a distinct divergence exists regarding the *Contradictory Statement* category. Taking Qwen3-4B as an example, *Contradictory Statement* errors escalate from 20.96% in intervention to 33.87% in counterfactual. This suggests that counterfactual tasks impose a higher demand on logical consistency. Instead of deriving the correct consequence mandated by the counterfactual assumption, LLMs frequently select options that directly conflict with the valid reasoning result. This reflects a deeper failure in faithfulness: LLMs are unable to faithfully execute the causal mechanisms described in the context to generate logically consistent outcomes under hypothetical premises.

4.4 Information Flow Analysis (RQ3)

To elucidate the findings from Section 4.3, this section analyzes the internal information flow of the LLMs utilizing *Saliency Scores* (Wang et al., 2023). By this means, we quantify the contribution of specific input segments to the final output, revealing the mechanistic drivers behind the divergent error patterns at each causal level.

Background on Information Flow Analysis. As established in Wang et al. (2023); Simonyan et al. (2013), the saliency score $I_l(i, j)$ quantifies the significance of the information flow from the j -th token to the i -th token at layer l of an LLM. Building on this, $S_{X \rightarrow Y}^{(l)}$ represents the aggregate significance of flow from a source segment X to a destination segment Y , calculated as the mean saliency across all token pairs connecting the two segments. By utilizing $S_{X \rightarrow Y}^{(l)}$, we can quantify the dependency strength of a specific component Y on a source component X within the prompt. Refer to Appendix C.4.1 or Wang et al. (2023) for details.

Experimental Setup. We randomly sample 500 instances from our dataset and manually annotate the evidence spans within the context that are essential for answering the questions. For each input

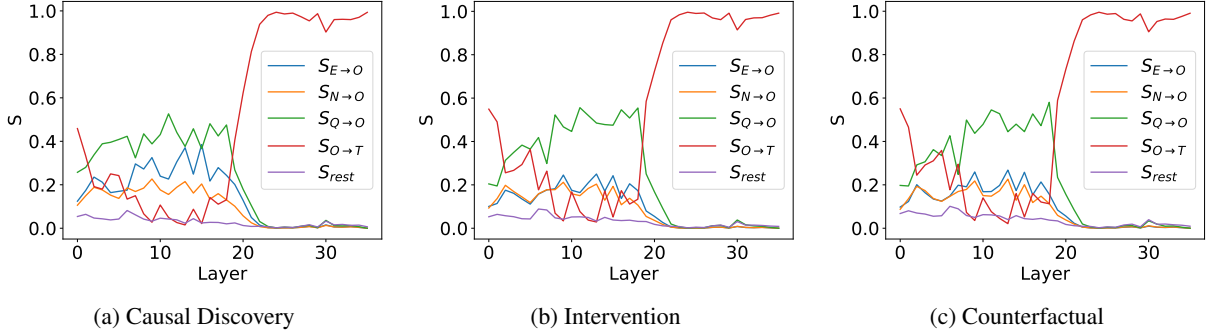


Figure 2: information flow significance at different layers.

prompt, our focus lies in five distinct components: (i) Evidence (E), the annotated causal clues; (ii) Non-evidence (N), contextual spans irrelevant to the question; (iii) Question (Q), the causal question; (iv) Selected Option (O), the option chosen by the corresponding LLM; and (v) Target (T), the final token position where the prediction is emitted (e.g., the colon ":" in the prompt "Pick one choice ... Answer:"). Based on these components, we analyze the following information flow metrics: $S_{E \rightarrow O}$, $S_{N \rightarrow O}$, $S_{Q \rightarrow O}$, $S_{O \rightarrow T}$, and S_{rest} that represents the mean significance among all remaining token pairs, excluding influences represented by the above metrics. Following Section 4.3, we focus on Qwen3-4b in the zero-shot setting, with more results detailed in Appendix C.4.2.

Mechanisms of Option Selection. We first investigate how models process the causal question and options to reach a decision. Figure 2 reveals that $S_{N \rightarrow O}$ and S_{rest} remain consistently low, indicating minimal interference from irrelevant background information. However, we observe distinct layer-wise dynamics: (1) Shallow Layers. Information flow patterns diverge by causal level. For causal discovery, both $S_{E \rightarrow O}$ and $S_{Q \rightarrow O}$ are high, suggesting the model actively utilizes both evidence and the question. In contrast, for intervention and counterfactual, while $S_{Q \rightarrow O}$ remains significant, $S_{E \rightarrow O}$ is notably suppressed. (2) Deep Layers. $S_{O \rightarrow T}$ becomes the dominant flow across all levels. Based on these observations, we hypothesize that in shallow layers, LLMs performing causal discovery integrate external evidence (E) with the question (Q) to anchor their selection. Conversely, at intervention and counterfactual levels, LLMs appear to rely primarily on the question to guide reasoning, underutilizing the specific evidence in favor of internal world knowledge. Subsequently, across all tasks, deep layers function to

Metric	Discovery		Intervention		Counterfactual	
	Cor.	Err.	Cor.	Err.	Cor.	Err.
$S_{E \rightarrow O}$	0.1690	0.1247	0.1144	0.0936	0.1095	0.0988
$S_{N \rightarrow O}$	0.0945	0.1262	0.0858	0.0858	0.0805	0.0884
$S_{Q \rightarrow O}$	0.2508	0.2163	0.2657	0.2378	0.2666	0.2281
$S_{O \rightarrow T}$	0.4685	0.4941	0.5039	0.5457	0.5071	0.5452
S_{rest}	0.0173	0.0387	0.0302	0.0371	0.0363	0.0395

Table 6: Layer-averaged significance scores for **Correct** vs. **Erroneous** predictions across causal levels.

propagate information from the identified option (O) to generate the final prediction (T). We conducted an additional check via attention masking to verify these causal contribution of evidence, the details of which are provided in Appendix C.4.3.

Error Analysis via Information Flow. We validate our error analysis (i.e., Section 4.3) by comparing information flow dynamics in correct vs. incorrect prediction settings. Specifically, Table 6 reports the layer-averaged values for the five significance metrics (i.e., averaged $S_{X \rightarrow Y}$). In causal discovery, erroneous predictions are characterized by a significant decrease in evidence utilization ($S_{E \rightarrow O} : 0.1690 \rightarrow 0.1247$) and a concurrent increase in noise sensitivity ($S_{N \rightarrow O} : 0.0945 \rightarrow 0.1262$). This provides empirical support for our observations in Section 4.3. In contrast, intervention and counterfactual tasks exhibit negligible reliance on either evidence or non-evidence (i.e., low $S_{E \rightarrow O}$ and $S_{N \rightarrow O}$), irrespective of prediction correctness. This context-agnostic behavior main explain the high incidence of hallucinated and contradictory responses (i.e., Unfounded and Contradictory Statements), as LLMs fail to ground their reasoning in the provided context. Furthermore, across all causal levels, incorrect predictions exhibit a lower $S_{Q \rightarrow O}$ compared to correct ones. It implies that LLM errors may also be associated

with an insufficient utilization of the causal question to guide the reasoning process.

Discussion. Our experimental findings reveal that LLMs appear to lack robust intrinsic capabilities for contextual causal reasoning, even at the architectural level. While the results in Table 4 demonstrate that advanced prompting schemes can yield marginal performance gains, significant room for improvement remains. The bottlenecks revealed by our information flow analysis, specifically the detachment from context and weak evidence grounding, suggest that Supervised Fine-Tuning or Reinforcement Learning may be necessary avenues for advancement. These training-based approaches could effectively reshape the LLM’s internal information flow, teaching it to prioritize contextual information, accurately localize evidence spans, and perform rigorous reasoning. Considering the high computational cost of model training, we devise a lightweight experiment to validate this hypothesis. By explicitly incorporating the evidence spans as a distinct component within the input prompt, we artificially enhance the significance of information flow from the evidence segment. The results indicate a consistent improvement in task performance across all levels (+3% for causal discovery, +2.5% for intervention, and +4.6% for counterfactual.), preliminarily confirming that guiding the LLM to attend more heavily to contextual evidence holds a potential as a viable pathway for enhancing contextual causal reasoning.

5 Conclusion

We present the first rigorous benchmark to systematically evaluate LLMs’ contextual causal reasoning capabilities under a unified context setting, traversing the full hierarchy of the causal ladder. Our study not only reveals significant limitations in LLM proficiency but also provides a deep analysis of these failures via error patterns and internal information flow. We highlight that LLMs appear to lack robust intrinsic capabilities for context-based causal reasoning. They exhibit high susceptibility to irrelevant contextual noise in Discovery level and a profound lack of faithfulness to the context in higher-level reasoning. This points toward the necessity of targeted training interventions to equip LLMs with advanced causal reasoning capabilities.

Limitations

Generalization of Mechanistic Findings to Black-box LLMs. Our mechanistic analysis of information flow is conducted on two distinct open-source architectures (Qwen and LLaMA). Furthermore, to assess the impact of scaling, we extend our experiments across varying parameter sizes within the Qwen family. However, due to the inaccessible nature of internal states in closed-source models such as the Gemini and GPT series, we are unable to perform equivalent information flow tracing on these systems. Consequently, while our findings are consistent across multiple open-weight architectures, the extent to which these mechanistic conclusions generalize to proprietary black-box LLMs remains to be verified.

Risk of Data Contamination. Our benchmark is derived from four open-source datasets. While this construction methodology aligns with established literature (Chi et al., 2024; Si et al., 2024), it introduces the potential risk of data contamination, as LLMs may have been pre-trained on these public corpora. We explicitly address this concern during data construction by implementing a rigorous Data De-contamination protocol (detailed in Section 3.2), which filters out samples that models can solve without context. However, given the opacity of pre-training data for LLMs, we cannot guarantee the complete elimination of all contamination risks.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative, (AISG Award No: AISG-NMLP-2024-002) and the National Natural Science Foundation of China (No. U25B201508, No. 62272330, and No.U24A20328). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

- Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.
- Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward

- mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.
- 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.
- Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86.
- Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. 2024. Causal evaluation of language models. *arXiv preprint arXiv:2405.00622*.
- Qing Cheng, Zefan Zeng, Xingchen Hu, Yuehang Si, and Zhong Liu. 2025a. A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *ACM Computing Surveys*, 58(3):1–37.
- Qing Cheng, Zefan Zeng, Xingchen Hu, Yuehang Si, and Zhong Liu. 2025b. A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *ACM Comput. Surv.*, 58(3).
- 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? *Advances in Neural Information Processing Systems*, 37:96640–96670.
- Dhairya Dalal, Paul Buitelaar, and Mihael Arcan. 2023. CALM-bench: A multi-task benchmark for evaluating causality-aware language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 296–311, Dubrovnik, Croatia. Association for Computational Linguistics.
- Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, and 1 others. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.
- Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA. Association for Computational Linguistics.
- Jörg Frohberg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, Marseille, France. European Language Resources Association.
- Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Xu Guo and Yiqiang Chen. 2024. Generative ai for synthetic data generation: Methods, challenges and the future. *arXiv preprint arXiv:2403.04190*.
- Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. 2023. Wikiwhy: Answering and explaining cause-and-effect questions. In *The Eleventh International Conference on Learning Representations*.
- 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.
- Abhinav Joshi, Ashutosh Modi, and 1 others. 2024. Cold: Causal reasoning in closed daily activities. *Advances in Neural Information Processing Systems*, 37:5145–5187.
- Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*.
- Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022. Meci: A multilingual dataset for event causality identification. In *Proceedings of the 29th international conference on computational linguistics*, pages 2346–2356.
- Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Uriel Anderson Lasheras and Vladia Pinheiro. 2025. CaLQuest.PT: Towards the collection and evaluation of natural causal ladder questions in Portuguese for AI agents. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 325–343, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.
- Haoyan Luo and Lucia Specia. 2024. From understanding to utilization: A survey on explainability for large language models. *arXiv preprint arXiv:2401.12874*.
- Jing Ma. 2025. Causal inference with large language model: A survey. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5886–5898, Albuquerque, New Mexico. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.
- Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models’ understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6277–6295, Bangkok, Thailand. Association for Computational Linguistics.
- Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia Passaro, Irene Sucameli, and Alessandro Lenci. 2025. ExpliCa: Evaluating explicit causal reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17335–17355, Vienna, Austria. Association for Computational Linguistics.
- Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, Bangkok, Thailand. Association for Computational Linguistics.
- Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.
- Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):3923.
- Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. 2023. CRAB: Assessing the strength of causal relationships between real-world events. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216, Singapore. Association for Computational Linguistics.
- Ryan Saklad, Aman Chadha, Oleg Pavlov, and Raha Moraffah. 2025. Can large language models infer causal relationships from real-world text? *arXiv preprint arXiv:2505.18931*.
- Jiasheng Si, Yibo Zhao, Yingjie Zhu, Haiyang Zhu, Wenpeng Lu, and Deyu Zhou. 2024. CHECKWHY: Causal fact verification via argument structure. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15636–15659, Bangkok, Thailand. Association for Computational Linguistics.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- Ya Su, Hu Zhang, Guangjun Zhang, Yujie Wang, Yue Fan, Ru Li, and Yuanlong Wang. 2025. Enhancing event causality identification with llm knowledge and concept-level event relations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7403–7414.
- Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Nelleke Oostdijk, Tommaso Caselli, Tadashi Nomoto, Onur Uca, Farhana Ferdousi Liza, and See Kiong Ng. 2023. Recess: Resource for extracting cause, effect, and signal spans. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–82.
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022a. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.
- Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.
- Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022b. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 (SIGHAN-10)*, pages 143–151.

Xing Wu, Shaoqi Peng, Jingwen Li, Jian Zhang, Qun Sun, Weimin Li, Qian Qian, Yue Liu, and Yike Guo. 2024. Causal inference in the medical domain: A survey. *Applied Intelligence*, 54(6):4911–4934.

Yongjie Xiao, Hongru Liang, Peixin Qin, Yao Zhang, and Wenqiang Lei. 2025. SCOP: Evaluating the comprehension process of large language models from a cognitive view. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17407–17431, Vienna, Austria. Association for Computational Linguistics.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5):1161–1186.

Linying Yang, Vik Shirvaikar, Oscar Clivio, and Fabian Falck. 2024. A critical review of causal reasoning benchmarks for large language models. *arXiv preprint arXiv:2407.08029*.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023. IfQA: A dataset for open-domain question answering under counterfactual presuppositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8276–8288, Singapore. Association for Computational Linguistics.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

A Comparison of Causal Reasoning Paradigms

To clarify the definition of Contextual Causal Reasoning proposed in this paper, we delineate its distinctions from the two existing paradigms: Commonsense Causal Reasoning and Formal Causal Reasoning. The core differences lie in the information source required for inference and the reasoning mechanism employed. Table 7 provides illustrative examples for each category.

Contextual Causal Reasoning. This paradigm is strictly evidence-driven. The causal conclusion is contingent upon specific, often idiosyncratic, details provided within a long-context narrative. As shown in Table 7, the model must comprehend the specific event sequence (e.g., Godflesh missing

the concert due to "permit issues") to answer the question. The reasoning process involves information retrieval, cross-sentence synthesis, and logical grounding within the provided document, rather than relying on external priors.

Commonsense Causal Reasoning. This paradigm is knowledge-driven. It relies on the LLM’s intrinsic parametric knowledge acquired during pre-training to resolve causal ambiguities. In the example, determining that "a blow to the head" causes unconsciousness requires general world knowledge about physiology and physical trauma, as this causal mechanism is not explicitly defined in the premise. The model acts as a knowledge base to judge plausibility.

Formal Causal Reasoning. This paradigm is rule-driven. It focuses on symbolic manipulation and probabilistic calculus within a closed, hypothetical system. The reasoning ignores real-world semantics (e.g., the definitions of "husband" or "alarm" are irrelevant) and strictly follows defined causal graphs or mathematical rules (e.g., Pearl’s do-calculus). As illustrated, the solution requires executing precise arithmetic operations based on conditional probabilities rather than interpreting natural language semantics.

B Benchmark Construction Details

Code and full datasets are available at <https://github.com/SCUNLP/METER>.

B.1 Data Preparation

We construct our initial data pool by sourcing from four established datasets, categorized into trigger-based and span-based types depending on their annotation granularity. Table 8 provides illustrative examples. For the trigger-based category, we utilize ESL (Caselli and Vossen, 2017), MAVEN-ERE (Wang et al., 2022b), and MECI (Lai et al., 2022), which primarily identify causality via specific event keywords. Specifically, the Event StoryLine Corpus (ESL) collects news articles regarding specific real-world events (e.g., natural disasters) to capture their narrative connections. To derive cause-effect pairs, we treat the keyword labeled as *rising action* as cause trigger and that labeled as *falling action* as effect trigger. MAVEN-ERE constructs a large-scale event relation dataset based on general-domain Wikipedia documents. Its causal annotation schema distinguishes between *CAUSE* (inevitable consequences) and *PRECONDITION* (nec-

Paradigm	Example Input	Model Reasoning Process & Answer
Contextual (<i>Evidence-driven</i>)	Context: In 1991, Godflesh embarked on their first North American tour. While most shows were successful, the schedule faced early setbacks. Specifically, Godflesh missed the first concert due to permit issues , though they made it to the second date in Brooklyn. Despite technical failures with their drum machine later, the rest of the tour concluded with positive reception. Question: Why did the band Godflesh miss their first concert?	Reasoning: I need to find the specific reason for missing the first concert in the provided text. Scanning the passage, I found the sentence: “ <i>Godflesh missed the first concert due to permit issues.</i> ” This explicitly states the cause. Answer: The band Godflesh ran into permit issues.
Commonsense (<i>Knowledge-driven</i>)	Premise: The man fell unconscious. What was the cause of this? Alternative 1: The assailant struck the man on the head. Alternative 2: The assailant took the man’s wallet.	Reasoning: I need to identify which event typically causes unconsciousness based on common sense. Alternative 1 describes a physical blow to the head, which causes trauma and is a medically plausible cause of unconsciousness. Alternative 2 describes a theft, which does not directly cause one to lose consciousness. Answer: Alternative 1.
Formal (<i>Rule-driven</i>)	Premise: Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships: Husband has a direct effect on wife and alarm clock. Wife has a direct effect on alarm clock. The overall probability of alarm set by husband is 77%. For husbands that don’t set the alarm, the probability of ringing alarm is 26%. For husbands that set the alarm, the probability of ringing alarm is 76%. Question: Is ringing alarm more likely than silent alarm overall?	Reasoning: 1. Define Variables: X =Husband, Y =Alarm. 2. Identify structure: $X \rightarrow Y$ and $X \rightarrow \text{Wife} \rightarrow Y$. 3. Apply Total Probability Theorem for $P(Y = 1)$: $P(Y) = P(Y X)P(X) + P(Y \neg X)P(\neg X)$ $= 0.76(0.77) + 0.26(1 - 0.77)$ $= 0.5852 + 0.0598 \approx 0.645$ 4. Compare with random chance (0.5): $0.645 > 0.5$, so ringing is more likely. Answer: Yes.

Table 7: Comparison of causal reasoning paradigms.

essary conditions). We exclusively filter for relations labeled as *CAUSE*, mapping the head event to the cause trigger and the tail event to the effect trigger. MECI establishes a multilingual benchmark based on Wikipedia articles covering five specific domains (e.g., aviation accidents). We utilize its English subset and derive cause-effect pairs by adhering to its directional annotation schema, which explicitly labels the source event as the *CAUSE* and the target event as the *EFFECT*. For the span-based category, we incorporate WIKIWHY (Ho et al., 2023), which captures causal logic through text segments spanning 11 diverse topics (e.g., Natural Sciences, History). We derive our data instances by aligning the annotated *cause* and *effect* spans based on their unique sample identifiers. In addition, the specific prompts utilized for event description expansion and the data de-contamination are provided in Table 17 and Table 18, respectively.

B.2 Data Generation

In this section, we provide implementation details for the data generation pipeline.

Question and Answer Generation. We employ a unified pipeline to generate question-answer pairs, ensuring consistency between the question and the ground truth. This process utilizes specific templates for question formulation, as listed in Table 9. And the specific prompts used are provided in Table 19. Finally, all generated questions are paraphrased by Gemini-2.5-pro to ensure natural, unambiguous phrasing.

For *Causal Discovery*, the templates consist of two categories: those inquiring about causes and those inquiring about effects. We instantiate these templates directly using the annotated cause or effect events, and use the corresponding effect or cause events as the answers. For each sample, we select one type of question to instantiate, ensuring a roughly equal distribution between cause-seeking and effect-seeking questions.

For *Intervention* and *Counterfactual*, we prompt Gemini-2.5-Pro to generate hypothetical conditions to fill the templates. Regarding intervention, the model generates specific, reasonable measures that do not contradict the existing background context.

Type: Trigger-based Dataset (e.g., MECI)

Context: The engine thrust reverser doors deployed, and the pilot decreased flaps from 40° to 15°. The landing gear remained locked in the down position. Six seconds before impact, when the aircraft was 4000 feet from the runway threshold, the aircraft climbed, then banked steeply to the left from a height of 300 to 400 feet, and crashed to the left of the runway. The aircraft was **destroyed** by **impact** and fire. The crash investigation was conducted by the Aviation Safety Investigation Division of Transport Canada...

Cause Trigger: impact

Effect Trigger: destroyed

Type: Span-based Dataset (e.g., WIKIWHY)

Context: Even though most ornithocheiromorphs didn't have a cranial crest like the closely related pteranodontids, there were some exceptions, this included *Caulkicephalus* and *Ludodactylus*. **Caulkicephalus had a rounded snout**, very similar to that of *Ornithocheirus* and *Anhanguera*, and therefore **it is placed within either Anhangueridae or Ornithocheiridae**, depending on the author. *Caulkicephalus* was also a large pterosaur, with wingspan estimates of around 5 meters (16 ft).

Cause Span: *Caulkicephalus* has a rounded snout.

Effect Span: *Caulkicephalus* is placed within either *Anhangueridae* or *Ornithocheiridae*.

Table 8: Data examples from the source datasets.

Causal Level	Question Template	
Causal Discovery	<i>Asking for Cause</i> <ul style="list-style-type: none">• Why does {EVENT} happen?• What causes {EVENT}?• What is the reason behind {EVENT}?• What leads to {EVENT}?	<i>Asking for Effect</i> <ul style="list-style-type: none">• What is the result of {EVENT}?• What does {EVENT} lead to?• What is the effect of {EVENT}?• What effect does {EVENT} have?
	Intervention	What will happen if {INTERVENTION}? What change will occur when {INTERVENTION CONDITION}? What will be the effect if {INTERVENTION CONDITION}? What will be the consequences of {INTERVENTION CONDITION}?
Counterfactual	What would happen if {COUNTERFACTUAL CONDITION}? If {COUNTERFACTUAL CONDITION}, what change would occur? Had {COUNTERFACTUAL CONDITION}, how would the situation be different? Assuming {COUNTERFACTUAL CONDITION}, what would be affected? What would be different if {COUNTERFACTUAL CONDITION}?	

Table 9: Templates used for generating questions across three causal reasoning levels.

Regarding counterfactual, the model modifies key attributes of the cause event (e.g., timing, participants or specific actions) to construct alternative scenarios. The corresponding answers are subsequently generated by the model, conditioned on the ground-truth causal events.

Distractor Generation. To construct effective distractors, we provide Gemini-2.5-Pro with specific definitions and formatting constraints for each distractor category. During generation, we enforce a rationale-augmented approach: the model is required to explicitly explain *why* a generated option is incorrect and *how* it satisfies the criteria of the assigned distractor type. This facilitates the generation of distractors that are unambiguously incorrect

and strictly adhere to the defined category constraints. The specific prompts used are provided in Table 20.

B.3 Human Verification Agreement

Annotator Recruitment. To handle the complexity of causal reasoning tasks, we recruited 9 undergraduate students with a solid background in Natural Language Processing (NLP). All annotators possess prior experience in linguistic analysis or logical reasoning tasks. Before the annotation commenced, all annotators were provided with comprehensive guidelines and reference examples to align their understanding of the inclusion criteria and error categories.

Model & Method	Causal Discovery (%)				Intervention (%)				Counterfactual (%)			
	Con	Unf	Ctr	Rev	Con	Unf	Ctr	Rev	Con	Unf	Ctr	Rev
Gemini3-Flash												
Zero-shot CoT	61.93	24.77	2.42	10.88	39.27	44.66	9.20	6.87	34.28	38.55	23.38	3.79
Few-shot	61.87	25.42	3.34	9.36	38.81	39.05	13.63	8.51	34.46	43.05	4.34	5.22
Few-shot CoT	61.45	27.11	2.71	8.73	43.41	39.02	8.16	9.41	38.15	40.00	17.73	4.12
Gemini3-Pro												
Zero-shot	57.78	23.70	8.15	10.37	35.97	35.34	14.10	14.58	33.56	42.78	17.65	6.02
Few-shot	54.65	27.52	6.98	10.85	36.07	40.98	12.02	10.93	29.49	48.01	17.15	5.35
Qwen3-4B												
Zero-shot CoT	49.64	24.72	5.70	19.94	24.77	38.08	20.89	14.52	20.78	40.30	28.53	7.52
Few-shot	50.75	26.03	7.12	16.10	26.51	38.08	20.89	14.52	21.96	37.31	33.21	7.52
Few-shot CoT	48.30	24.26	8.94	18.51	25.18	35.19	23.33	16.30	21.34	37.70	34.57	6.40

Table 10: Detailed error distribution analysis for *Gemini3-Flash*, *Gemini3-Pro*, and *Qwen3-4B* under various prompting strategies. The values represent the percentage (%) of each error type within the incorrect predictions. **Con**: Contextual (Irrelevant Fact), **Unf**: Unfounded, **Ctr**: Contradictory, **Rev**: Causal Reversal.

Quality Control. We implemented a multi-stage validation process.

- **Phase 1: Raw Data Verification.** Three annotators independently reviewed the extracted event pairs. To ensure high precision, any instance rejected by at least one annotator due to ambiguity or incompleteness was immediately discarded.
- **Phase 2: Generated Data Validation.** The validation of LLM-generated content involved two distinct steps: *Editing* and *Filtering*.
 1. **Manual Editing.** Given each instance, two annotators are required to perform manual editing of the generated question, answer, and distractors, and to resolve any errors identified by a third annotator. The final result is chosen by voting from all three annotators.
 2. **Manual Filtering.** We applied a strict quality control procedure by filtering each edited instance. Three annotators independently inspected each entry against the verification standards. Any sample failing to meet these standards was discarded. The final decision to retain or reject a sample was reached through a majority vote among the three annotators.

C Experiment Details

C.1 Human Baseline

To establish a human performance reference, we randomly sample 100 instances from our dataset. We recruit five undergraduate students with NLP backgrounds to independently answer these questions based strictly on the provided context.

C.2 Methods & Implementation Details

We evaluate closed-source models via their official APIs. For open-source models, we deploy them using the **vLLM** framework on a server equipped with four NVIDIA A100 GPUs. To ensure controllable results and minimize generation randomness, we set the decoding temperature to 0 for all models, consistent with the configuration in Xiao et al. (2025). Regarding the prompting schemes, we implement specific protocols to standardize the evaluation. For *Zero-shot CoT*, we append the standard trigger phrase “Let’s think step by step” to the input prompt to elicit reasoning. For *Few-shot CoT*, we utilize demonstrations where the reasoning process is manually written by human annotators. We employ a targeted prompt strategy in the Causal Discovery task for *Few-shot* and *Few-shot CoT* settings: we provide distinct sets of demonstrations for questions inquiring about causes and those inquiring about effects, respectively. This ensures that the few-shot examples align strictly with the direction of the causal inquiry.

C.3 Error Analysis

In this section, we extend the failure mode analysis presented in Section 4.3 to cover a broader spectrum of models and prompting strategies. Specifically, we report the error distribution for *Gemini3-Pro*, *Gemini3-Flash*, and *Qwen3-4B* across four prompt settings, as illustrated in Table 10. We observe that while the exact error rates fluctuate depending on the model architecture and prompting paradigm, the overall failure patterns remain

Metric	Causal Discovery		Intervention		Counterfactual	
	Correct	Error	Correct	Error	Correct	Error
Qwen3-8B						
$S_{E \rightarrow O}$	0.1682	0.1058	0.1056	0.0853	0.1095	0.0845
$S_{N \rightarrow O}$	0.0856	0.1006	0.0785	0.0801	0.0798	0.0784
$S_{Q \rightarrow O}$	0.2048	0.2008	0.2548	0.2230	0.2448	0.2158
$S_{O \rightarrow T}$	0.5243	0.5564	0.5329	0.5723	0.5354	0.5801
S_{rest}	0.0171	0.0364	0.0283	0.0393	0.0306	0.0412
Llama-3.2-3B						
$S_{E \rightarrow O}$	0.1086	0.0679	0.0595	0.0493	0.0608	0.0477
$S_{N \rightarrow O}$	0.0592	0.0696	0.0504	0.0422	0.0527	0.0419
$S_{Q \rightarrow O}$	0.1471	0.1357	0.1660	0.1263	0.1537	0.1056
$S_{O \rightarrow T}$	0.6614	0.6966	0.6944	0.7376	0.6970	0.7285
S_{rest}	0.0238	0.0302	0.0297	0.0446	0.0357	0.0494

Table 11: Comparison of layer-averaged significance scores between correct and incorrect predictions for *Qwen3-8B* and *Llama-3.2-3B*.

largely consistent with the conclusions drawn in Section 4.3.

C.4 Information Flow

In this section, we provide supplementary details regarding the information flow analysis.

C.4.1 Saliency Scores

We adopt the methodology proposed by (Wang et al., 2023), which leverages the saliency technique (Simonyan et al., 2013) to quantify token interactions. Specifically, the saliency score for each element of the attention matrix is computed as:

$$I_l = \sum_h \left| A_{h,l}^\top \frac{\partial \mathcal{L}(x)}{\partial A_{h,l}} \right|, \quad (1)$$

where $A_{h,l}$ is the attention matrix of the h -th head in the l -th layer, and $\mathcal{L}(x)$ is the loss function given the input x , which is the cross-entropy loss in our task. $I_l(i, j)$ represents the significance of the information flow from the j -th token to the i -th token. For more details, see (Wang et al., 2023).

To quantify the interactions between segments, we define the *mean significance of information flow* $S_{X \rightarrow Y}$ from a source segment X to a destination segment Y as:

$$S_{X \rightarrow Y} = \frac{\sum_{(i,j) \in \mathcal{C}_{X \rightarrow Y}} I_l(i, j)}{|\mathcal{C}_{XY}|}, \quad (2)$$

where $\mathcal{C}_{XY} = \{(i, j) : i \in Y, j \in X\}$ denotes the set of token pairs representing the flow from X to Y , and $|\mathcal{C}_{XY}|$ is the cardinality of this set. Based

on this definition, we introduce five quantitative metrics:

$S_{E \rightarrow O}$: mean significance from evidence to the selected option.

$S_{N \rightarrow O}$: mean significance from non-evidence to the selected option.

$S_{Q \rightarrow O}$: mean significance from question to the selected option.

$S_{O \rightarrow T}$: mean significance from the selected option to the target.

S_{rest} : mean significance among all remaining token pairs, excluding influences represented by the above metrics

$$S_{rest} = \frac{\sum_{(i,j) \in \mathcal{C}_{rest}} I_l(i, j)}{|\mathcal{C}_{rest}|}, \quad (3)$$

where $\mathcal{C}_{rest} = \{(i, j) : j < i\} - (\mathcal{C}_{EO} \cup \mathcal{C}_{NO} \cup \mathcal{C}_{QO} \cup \mathcal{C}_{OT})$. where $\mathcal{C}_{rest} = \{(x, y) : x \leq y\} - (\mathcal{C}_{EO} \cup \mathcal{C}_{NO} \cup \mathcal{C}_{QO} \cup \mathcal{C}_{OT})$. Here, \mathcal{C}_{XY} denotes the set of pairs in $X \times Y$.

C.4.2 More Results

We further examine the information flow dynamics of *Llama-3.2-3B* and *Qwen3-8B*, adhering to the same experimental setup described in Section 4.4. Specifically, we track the changes in information flow across layers, comparing the information flow differences between correct and incorrect predictions. As illustrated in Figure 3 and Table 11, the observed phenomena are highly consistent with the findings reported in Section 4.4.

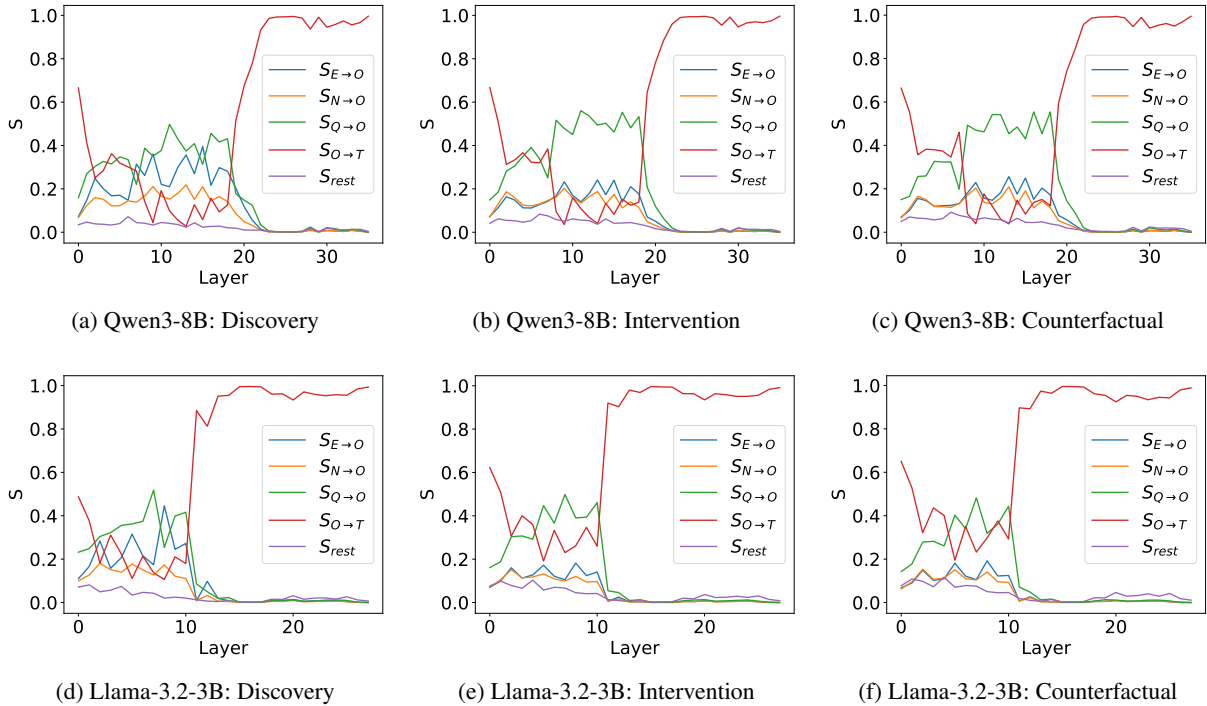


Figure 3: Layer-wise information flow dynamics for *Qwen3-8B* (Top Row) and *Llama-3.2-3B* (Bottom Row) across three causal reasoning levels. The plots illustrate the aggregation trends of saliency scores, contrasting the patterns between correct and incorrect predictions.

C.4.3 Causal Verification via Attention Masking

To further verify the causal contribution of contextual evidence and validate the layer-wise aggregation patterns observed in our information flow analysis, we conducted an attention masking experiment on *Qwen3-4B*.

Methodology. We block specific information flows by manipulating the attention mechanism. To isolate the destination tokens from the source tokens, we manipulate the attention matrix A . Specifically, we set $A_l(p, i) = 0$ (where $i < p$) in the attention matrix of the l -th layer, where p represents the index of the destination tokens (e.g., Options) and i represents the index of the preceding source tokens (e.g., Evidence). Consequently, in the l -th layer, the destination tokens cannot access information from the source tokens.

Experimental Setup. Based on our observation that evidence aggregation predominantly occurs in shallow layers (stabilizing around Layer 24), we designed two distinct masking settings to verify this temporal dynamics:

- **Shallow Masking:** Blocking Evidence \rightarrow Option ($E \rightarrow O$) flow in **Layers 1–24**.
- **Deep Masking:** Blocking Evidence \rightarrow Option

($E \rightarrow O$) flow in **Layers 25–End**.

Results. The results are illustrated in Figure 4.

- **Shallow Layers (1–24):** Masking the evidence flow in shallow layers leads to a significant accuracy drop in Causal Discovery (from 0.827 to 0.579). In contrast, the performance for Intervention and Counterfactual tasks remains largely unaffected (e.g., Intervention holds steady at ~ 0.53). This causally confirms two findings: (1) Causal Discovery heavily relies on evidence aggregation in shallow layers; (2) Higher-level tasks underutilize the provided context, relying instead on internal priors.
- **Deep Layers (25–End):** Masking the evidence flow in deep layers results in **no performance change** across all three tasks (accuracy remains identical to the baseline). This validates our hypothesis that the aggregation of evidentiary information is completed in the shallow layers, and deep layers do not revisit the evidence tokens for reasoning.

D Additional Experiment Results

In this section, we provide detailed results for the hypothesis validation experiment discussed in

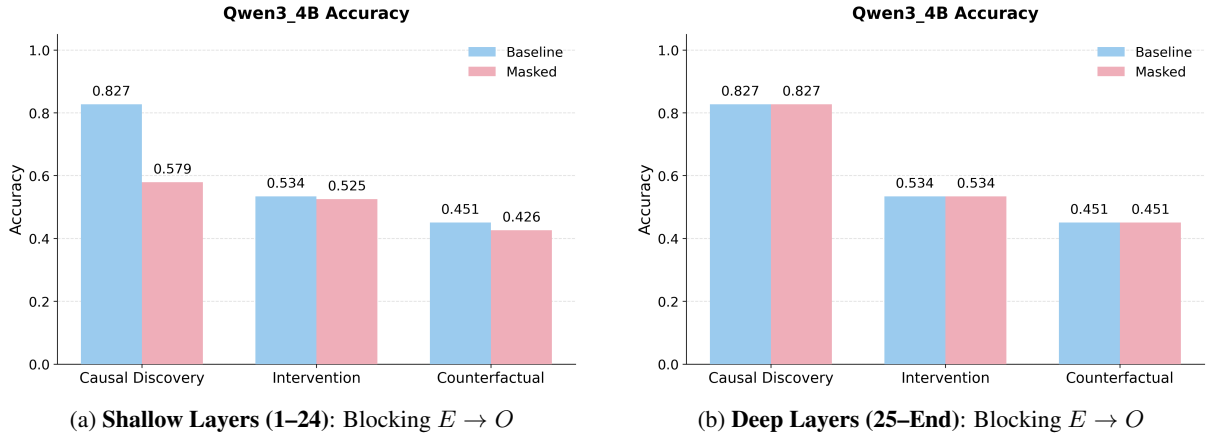


Figure 4: Results of the Attention Masking experiment on *Qwen3-4B*. (a) Masking evidence-to-option attention in shallow layers causes a significant drop in Causal Discovery accuracy, confirming reliance on evidence aggregation. (b) Masking the same flow in deep layers results in no performance change, indicating evidence processing is completed in early layers.

the Section 4.4. To verify whether strengthening evidence grounding improves contextual causal reasoning, we devised a lightweight intervention: we explicitly appended the ground-truth evidence spans to the input prompt, thereby artificially enhancing their visibility to the model.

D.1 Information Flow Analysis.

We computed the information flow metrics following the setup described in Section 4.4. We compared the saliency scores under the original setting (*No Evidence*) versus the enhanced setting (*Evidence*). As detailed in Table 12, the explicit inclusion of evidence results in a consistent increase in the information flow from the evidence segment to the selected option (denoted as $S_{E \rightarrow O}$) across various models and causal levels. For instance, in the causal discovery task, *Llama-3.2-3B* exhibits a rise in $S_{E \rightarrow O}$ from 0.0999 to 0.1190. In the intervention task, *Qwen3-8B* shows an increase from 0.0902 to 0.1006. Similarly, for counterfactual reasoning, *Qwen3-4B* demonstrates a growth from 0.1003 to 0.1066. These variations indicate that providing explicit evidence effectively amplifies the propagation of information from the supporting facts to the model’s final decision.

D.2 Performance Improvement.

We further evaluated the impact of this enhanced information flow on reasoning accuracy. As shown in Table 13, explicitly providing evidence yields consistent performance gains. For instance, *Qwen3-8B* achieves approximate improvements of +3.0% in Causal Discovery, +3.3% in Intervention, and

+4.8% in Counterfactual. These empirical results confirm that mitigating the bottleneck of weak context utilization by reinforcing the information flow from evidence is a viable pathway for enhancing contextual causal reasoning.

D.3 Error Analysis

We further tracked the fluctuation in error counts to quantify the impact of incorporating evidence. As illustrated in Table 14, providing explicit evidence leads to a reduction in *Unfounded*, *Irrelevant Fact*, and *Contradictory* errors across most models. Specifically, in the intervention task, *Qwen3-4B* reduced Unfounded errors from 95 to 80. Similarly, in the causal discovery task, *Qwen3-8B* saw a significant drop in Irrelevant Fact errors from 44 to 29. These declines confirm that evidence acts as a rigid constraint, effectively grounding the model’s generation to the provided context and minimizing hallucinations.

However, we observed a counter-intuitive side effect: a consistent rise in *Causal Reversal* errors. For instance, *Qwen3-4B* saw an increase in reversal errors from 22 to 35 in the Intervention task, while *Llama-3.2-3B* exhibited a similar trend in Causal Discovery (rising from 37 to 46). We speculate that this may occur because evidence spans reinforce the model’s focus on specific keywords or events, but not their logical order. Since *Causal Reversal* options typically contain the same entities as the evidence (just in the wrong direction), the model might be biased towards these options due to high text similarity.

Metric	Causal Discovery		Intervention		Counterfactual	
	Evidence	No Evidence	Evidence	No Evidence	Evidence	No Evidence
Qwen3-4B						
$S_{E \rightarrow O}$	0.1481	0.1422	0.1080	0.0968	0.1066	0.1003
$S_{N \rightarrow O}$	0.1034	0.0957	0.0852	0.0863	0.0916	0.0872
$S_{Q \rightarrow O}$	0.2066	0.2312	0.2204	0.2425	0.2145	0.2336
$S_{O \rightarrow T}$	0.5132	0.5031	0.5558	0.5421	0.5494	0.5401
S_{rest}	0.0288	0.0279	0.0307	0.0323	0.0379	0.0387
Llama-3.2-3B						
$S_{E \rightarrow O}$	0.1190	0.0999	0.0639	0.0536	0.0641	0.0520
$S_{N \rightarrow O}$	0.0573	0.0592	0.0447	0.0457	0.0461	0.0468
$S_{Q \rightarrow O}$	0.1234	0.1444	0.1310	0.1433	0.1177	0.1328
$S_{O \rightarrow T}$	0.6754	0.6700	0.7287	0.7212	0.7282	0.7260
S_{rest}	0.0249	0.0265	0.0317	0.0362	0.0439	0.0424
Qwen3-8B						
$S_{E \rightarrow O}$	0.1587	0.1412	0.1006	0.0902	0.1030	0.0914
$S_{N \rightarrow O}$	0.0882	0.0878	0.0786	0.0802	0.0815	0.0788
$S_{Q \rightarrow O}$	0.1932	0.2031	0.2135	0.2310	0.2036	0.2233
$S_{O \rightarrow T}$	0.5374	0.5438	0.5780	0.5655	0.5754	0.5687
S_{rest}	0.0225	0.0241	0.0293	0.0331	0.0365	0.0378

Table 12: Impact of explicitly incorporating evidence spans on information flow dynamics. Values represent the saliency scores under two settings: with explicitly provided evidence (**Evidence**) and without it (**No Evidence**).

Model	Causal Discovery		Intervention		Counterfactual	
	Evidence	No Evidence	Evidence	No Evidence	Evidence	No Evidence
Qwen3-4B	84.80	82.67	55.20	53.40	46.40	45.06
Qwen3-8B	88.00	85.40	65.85	63.71	56.40	53.80
Llama-3.2-3B	75.00	72.60	57.83	57.40	44.60	42.03

Table 13: Performance comparison (Accuracy) between the original setting (**No Evidence**) and the setting with explicitly incorporated evidence spans (**Evidence**).

E Data Example and Case Studies

E.1 Data Example

Table 15 presents representative samples from our dataset. Each entry consists of a context paragraph, followed by three types of causal questions (causal discovery, intervention, and counterfactual), along with their corresponding candidate options and correct answers.

E.2 Case Studies

Table 16 presents a failure instance of *Gemini3-Flash* (Few-shot) in the causal discovery task. The model was prompted to identify the reason why a referee stopped a boxing match based on a given context. The correct answer is (A), which identifies

the root cause: the challenger’s inability to recover after being knocked down. However, *Gemini3-Flash* incorrectly selected the distractor (B). This distractor, while factually present in the text, misattributes the cause to the procedural step of the referee giving a standing count, rather than the underlying physical condition of the fighter. We speculate this error stems from the model’s focus on explicitly mentioned actions in the text. As indicated in its reasoning, the model found Option B to be more "precise about the immediate officiating action." This suggests that the LLM might be prioritizing surface-level textual cues and procedural descriptions over inferring the deeper causal relationship, such as the fighter’s actual physical state

Error Type	Qwen3-4B			Qwen3-8B			Llama-3.2-3B		
	Disc.	Inter.	Count.	Disc.	Inter.	Count.	Disc.	Inter.	Count.
Irrelevant Fact	48→ 35	69→ 67	91→ 87	44→ 29	76→ 72	57→ 56	46→ 36	27→ 25	47→ 36
Unfounded	26→ 25	95→ 80	75→ 73	24→ 23	61→ 52	77→ 66	34→ 28	88→ 76	107→ 95
Contradictory	7→ 5	47→ 42	98→ 95	5→ 4	31→ 25	89→ 85	20→ 15	72→ 73	95→ 97
Causal Reversal	6→ 11	22→ 35	11→ 23	0→ 4	13→ 22	8→ 11	37→ 46	26→ 37	41→ 49

Table 14: Shift in error distribution after incorporating evidence spans (format: No Evidence → **With Evidence**). **Blue values** indicate a reduction in errors (improvement), while **Red values** indicate an increase (degradation). Note the consistent reduction in Irrelevant and Unfounded errors versus the rise in Causal Reversal errors.

necessitated by the knockdown.

F Prompts

Detailed examples of the prompts used in both our dataset construction and experiments are provided in Tables 17 through 21.

Context

Henry VI's two envoys landed in Cyprus in April or May 1196. Aimery may have adopted the title of king around that time, because Pope Celestin styled him as king already in a letter in December 1196. In the same month, the Pope set up a Roman Catholic archdiocese in Nicosia with three suffragan bishops in Famagusta, Limassol and Paphos. The Greek Orthodox bishops were not expelled, but their property and income was seized by the new Catholic prelates.

Level 1: Causal Discovery

Question: Why did Aimery of Cyprus adopt the title of king around April or May of 1196?

Options:

- A. The confiscation of the Greek Orthodox bishops' property and income by Aimery's newly installed Catholic prelates later in 1196.
- B. He had recently married a Byzantine princess and needed a title that matched her royal rank.
- C. Aimery's assumption of the royal title led Pope Celestine to call him king in the December 1196 letter.
- D. Pope Celestine expressly refused to address Aimery as king in his December 1196 letter.
- E. Pope Celestine styled Aimery of Cyprus as king already in a letter in December 1196.

Correct Answer: E

Level 2: Intervention

Question: What will happen if the Pope decrees that the new Catholic prelates must share a portion of their new income with the formerly dispossessed Greek Orthodox bishops?

Options:

- A. The new Catholic prelates will use the mandated payments to pressure the Greek Orthodox bishops into publicly accepting their authority.
- B. Pope Celestin will issue another letter officially recognizing Aimery's kingship over Cyprus.
- C. The Greek Orthodox bishops' successful appeal to King Aimery will cause the Pope to issue the decree for income sharing.
- D. The Greek Orthodox bishops will be expelled from Cyprus to resolve the ongoing dispute over church income.
- E. The local Cypriot population will revolt in support of the Greek Orthodox bishops, demanding the full restoration of their property.

Correct Answer: A

Level 3: Counterfactual

Question: What would have happened if Pope Celestine had addressed Aimery only as 'Lord of Cyprus' in his December letter?

Options:

- A. Pope Celestine would not have established a Roman Catholic archdiocese in Nicosia.
- B. Aimery's failure to seize the property of the Greek Orthodox bishops would have caused the Pope to address him as 'Lord of Cyprus'.
- C. Aimery would have immediately sought a royal title from the Holy Roman Emperor, Henry VI.
- D. The establishment of the Roman Catholic archdiocese in Nicosia would have been postponed.
- E. Aimery would have been forced to offer greater concessions to the new Catholic archdiocese to secure his royal legitimacy.

Correct Answer: E

Table 15: Data samples from our dataset based on a single context.

Case Study: Causal Discovery with Gemini3-Flash (Few-shot)

Context: Wladimir Klitschko stops Tony Thompson in 6th round to retain world heavy-weight titles... The Ukrainian champion dropped Thompson again in the sixth, and referee Sam Williams gave the challenger a standing count before stopping the fight...

Question: What caused referee Sam Williams to stop the fight?

Options:

A (Correct - Root Cause): The fight was stopped due to the challenger's inability to recover after being knocked down by the champion's powerful punches in consecutive rounds.

B (Model's Choice - Procedural Distractor): The fight was stopped because the referee's decision to give a standing count determined that the challenger was no longer fit to continue.

(Options C, D, and E are omitted for brevity)

Model Reasoning:

"The referee stopped the fight after Wladimir Klitschko knocked Tony Thompson down in the sixth round and administered a standing count, indicating Thompson could no longer continue. Choice A is partially correct regarding the knockdowns but less precise about the immediate officiating action. Choice B accurately identifies the referee's assessment and the standing count as the direct cause for the stoppage."

Table 16: A failure case of *Gemini3-Flash*.

Event Description Expansion

You are an expert in causal event extraction. Given a paragraph P and two event triggers e1 and e2, your task is to identify and summarize the complete events they represent. The event e1 is the cause of e2.

Definitions:

– **Event Trigger:** A specific word (typically a verb or action-related noun) that serves as the core or anchor of an event.

Guidelines:

For each trigger, locate it in the paragraph and summarize its corresponding event. Each summary must be:

– **Accurate:** only what the trigger states, no extra causes/effects.

– **Concise:** no extra background or causal details.

– **Complete:** forms a minimal grammatical event (subject-verb or subject-verb-object).

– **One-to-one:** each trigger corresponds to exactly one event.

Paragraph: <context>

e1: <trigger1> **e2:** <trigger2>

Answer:

Table 17: The prompt used to expand event triggers into complete event descriptions.

Data De-contamination

You are an expert in causal inference. Your task is to analyze Event A and Event B provided below and determine whether there is a causal relationship between them. Respond with "Yes" if a causal relationship exists, and "No" otherwise.

Event A: <event1> **Event B:** <event2>

Answer:

Table 18: The prompt used to filter out cause-effect pairs that LLMs might have memorized.

Intervention Question and Answer Generation

You are an expert in causal inference, expert at crafting deep, multi-layered causal inquiries. You will be given:

Paragraph: a paragraph describing a causal scenario.

Cause and Effect Events: two events, which are causally connected.

Your task is to instantiate the template below to design a **intervention question** and give a correct answer based on the given cause and effect events.

The intervention question explores the future causal consequences of introducing a new and specific action (the “intervention”). Crucially, the intervention must be specific and explore a diverse state, not a vague or simple negation. You cannot modify or negate the fixed history of the scenario. The question must be in the future tense grammatically. The answer must be tightly grounded in the Paragraph, serving as a logical extension of the facts and dynamics provided. Ensure the answer is concise, accurate.

Question Template: <question template>

Paragraph: <context>

Cause Event: <cause>

Effect Event: <effect>

Answer:

Counterfactual Question and Answer Generation

You are an expert in causal inference, expert at crafting deep, multi-layered causal inquiries. You will be given:

Paragraph: a paragraph describing a causal scenario.

Cause and Effect Events: two events, which are causally connected.

Your task is to instantiate the template below to design a **counterfactual question** and give a correct answer based on the given cause and effect events.

The counterfactual question explores the causal consequences of a hypothetical change to a key condition or action from the past event. Crucially, this change should be specific and imaginative, not just a simple negation of the event. Consider altering the event’s timing, intensity, method, or substituting a key element. The answer should provide an insightful inference. Focus on the most significant shift the counterfactual change would have caused. Ensure the answer is concise, accurate and reasonably inferred from the paragraph.

Question Template: <question template>

Paragraph: <context>

Cause Event: <cause>

Effect Event: <effect>

Answer:

Table 19: Prompts used for generating Intervention and Counterfactual questions along with their answers.

Distractor Generation

You are an expert in constructing multiple-choice causal reasoning questions. You will be given:

- **Paragraph:** a paragraph describing a causal scenario.
- **Question:** a <question_type>.
- **Correct Option:** the correct option.

Your task is to generate four distinct and misleading incorrect options (distractors), each corresponding to a different fallacy type. For each generated distractor, you must also provide a brief rationale explaining why the option is incorrect yet sufficiently plausible to be misleading.

Distractor Types:

- **Contradictory Statement:** An option contains information conflicting with a specific fact stated in the paragraph or premises in the question.
- **Unfounded Statement:** An option contains information not stated in or inferred from the paragraph.
- **Causal Reversal:** An option inverts the cause-and-effect direction established in the paragraph.
- **Irrelevant Fact:** An option that is grounded in the paragraph but has no causal link to the question.

Important Rules:

- Avoid obviously absurd or trivial answers; all distractors should require careful reasoning to eliminate.
- Style and length of distractors should match the correct answer.

Output Format

- Only return a valid JSON object with the following structure:

```
{  
  "contradictory": {  
    "option_text": "<distractor_1>",  
    "reasoning": "<text>"  
  },  
  "unfounded": {  
    "option_text": "<distractor_2>",  
    "reasoning": "<text>"  
  },  
  "reversal": {  
    "option_text": "<distractor_3>",  
    "reasoning": "<text>"  
  },  
  "irrelevant": {  
    "option_text": "<distractor_4>",  
    "reasoning": "<text>"  
  }  
}
```

Paragraph: <context>

Question: <question>

Correct Option: <correct option>

Answer:

Table 20: Prompt used for generating distractors.

Prompting Schemes

Zero-shot: Pick one choice given the context to answer the question.

Context: $\langle Context \rangle$ **Question:** $\langle Question \rangle$ **Choices:** A. $\langle Choice A \rangle$ B. $\langle Choice B \rangle$ C. $\langle Choice C \rangle$ D. $\langle Choice D \rangle$ E. $\langle Choice E \rangle$ **Answer:**

Few-shot: Pick one choice given the context to answer the question. Example:

Context: The city library, usually open late, closed two hours early last month. For several days beforehand, the lights in the reading rooms kept flickering. In a council bulletin, officials explained that the building’s electrical system had problems, though some residents blamed budget cuts. **Question:** Why did the library close earlier than usual last month? **Choices:** A. The city held a council bulletin. B. The library’s electrical system was malfunctioning. C. Closing early caused the lights to flicker. D. Budget cuts reduced funding for cultural programs. E. The library lights flickered in the evenings. **Answer:** B

Context: $\langle Context \rangle$ **Question:** $\langle Question \rangle$ **Choices:** A. $\langle Choice A \rangle$ B. $\langle Choice B \rangle$ C. $\langle Choice C \rangle$ D. $\langle Choice D \rangle$ E. $\langle Choice E \rangle$ **Answer:**

Zero-shot CoT: Pick one choice given the context to answer the question.

Context: $\langle Context \rangle$ **Question:** $\langle Question \rangle$ **Choices:** A. $\langle Choice A \rangle$ B. $\langle Choice B \rangle$ C. $\langle Choice C \rangle$ D. $\langle Choice D \rangle$ E. $\langle Choice E \rangle$ Let’s think step by step.

Few-shot CoT: Pick one choice given the context to answer the question. Example:

Context: The city library, usually open late, closed two hours early last month. For several days beforehand, the lights in the reading rooms kept flickering. In a council bulletin, officials explained that the building’s electrical system had problems, though some residents blamed budget cuts. **Question:** Why did the library close earlier than usual last month? **Choices:** A. The city held a council bulletin. B. The library’s electrical system was malfunctioning. C. Closing early caused the lights to flicker. D. Budget cuts reduced funding for cultural programs. E. The library lights flickered in the evenings.

Answer: Officials stated the building’s electrical system had problems, which explains the early closure. A is irrelevant: the council bulletin is merely where the explanation appears, not a cause. B is correct: it aligns perfectly with the official explanation. C reverses causality: flickering happened beforehand and is due to electrical issues, not caused by closing early. D is unsupported conjecture (residents’ blame) and is outweighed by the official explanation. E describes a symptom/timing detail, not the reason for closing early. **Correct answer:** B

Context: $\langle Context \rangle$ **Question:** $\langle Question \rangle$ **Choices:** A. $\langle Choice A \rangle$ B. $\langle Choice B \rangle$ C. $\langle Choice C \rangle$ D. $\langle Choice D \rangle$ E. $\langle Choice E \rangle$ **Answer:**

Table 21: Prompting templates used in our experiments across four settings: Zero-shot, Few-shot, Zero-shot CoT, and Few-shot CoT.