

Contextual Causality with Large Language Models: A Survey

Yiheng Zhao^{1,†}, Jun Yan¹, Chengming Hu²

¹Concordia University, Montreal, Canada

²McGill University, Montreal, Canada

[†]Correspondence: yiheng.zhao@mail.concordia.ca

Abstract

Understanding contextual causality is critical for large language models (LLMs), as it enables them to accurately identify causal relations in specific situations and support more reliable decision-making. Despite its significance, a systematic exploration of contextual causality with LLMs is still lacking. To fill this gap, we present a comprehensive survey on this topic. In this survey, we first propose a taxonomy of contextual causality, consisting of semantic, intervention, and counterfactual causality, and characterize each category by its core causal question, required model capabilities, representative tasks, and practical uses in causality analysis. We then analyze existing studies and discuss their key limitations. Finally, we examine the gaps between current benchmarks and real-world needs and outline promising directions for future research. Our goal is to clarify the research landscape of contextual causality with LLMs, emphasize its importance, and highlight promising future directions.

1 Introduction

Human knowledge and human power meet in one; for where the cause is not known the effect cannot be produced.
— Francis Bacon

Causality, an important aspect of intelligence, can be broadly divided into non-contextual and contextual forms (Lake et al., 2017; Halpern, 2016). Non-contextual causality refers to general causal relations between variables, such as smoking increasing the risk of lung cancer. In contrast, contextual causality concerns particular causal relations grounded in concrete situations, such as whether a specific patient’s lung cancer is caused by long-term smoking. In natural language settings, contextual causality is grounded in a given natural language situation, involving causal relations either among events or between actions and outcomes.

Understanding such causality is essential for large language models (LLMs), as it enables them to accurately identify causal relations and better estimate the outcomes of actions, thereby supporting more reliable decision-making (Drury et al., 2022; Wang, 2024; Yang et al., 2022; Halpern, 2016; Lake et al., 2017). This is especially important in high-stakes domains (Richens et al., 2020; Wu et al., 2024). For example, in personalized medical diagnosis, LLMs need to identify the causal relations underlying a patient’s symptoms, medical history, and clinical records to identify factors driving the current condition, anticipate the effects of different treatments, and support more reliable decisions.

Despite its importance, a systematic overview of contextual causality with LLMs remains lacking. Although several surveys have recently discussed the role of LLMs in causality, most of them primarily focus on non-contextual causality, including causal discovery and causal inference with LLMs (Wan et al., 2024; Liu et al., 2025; Ma, 2025; Yu et al., 2025; Zhou et al., 2025; Bazgir et al., 2025; Li et al., 2025b). There are also task-specific surveys on problems related to contextual causality, such as counterfactual generation (CG) with LLMs and event causality identification (ECI) (Wang et al., 2024b; Cheng et al., 2025). However, these surveys cover only a limited portion of the broader landscape of contextual causality. In particular, the ECI survey reviews the development of ECI and includes only limited coverage of recent studies on ECI with LLMs. In addition, some surveys discuss broader thematic topics such as commonsense causality, causal uncertainty, and benchmark analysis (Cui et al., 2024, 2025; Yang et al., 2024b). Therefore, despite its importance, a systematic overview of contextual causality with LLMs remains lacking. To fill this gap, we conduct a comprehensive survey on this topic.

In this survey, we first introduce a taxonomy of contextual causality, motivated by causal analysis

Survey	Contextual Causality			Non-Contextual Causality	
	Semantic	Intervention	Counterfactual	Causal Discovery	Causal Inference
(Wan et al., 2024)	✗	✗	✗	✓	✗
(Liu et al., 2025)	✗	✗	✗	✓	✓
(Ma, 2025)	✗	✗	✗	✗	✓
(Yu et al., 2025)	✗	✗	✗	✓	✓
(Zhou et al., 2025)	⦿	✗	✗	⦿	⦿
(Bazgir et al., 2025)	✗	✗	✗	✓	✓
(Li et al., 2025b)	✗	✗	✗	⦿	⦿
(Wang et al., 2024b)	✗	✗	✓ _{CG}	✗	✗
(Cheng et al., 2025)	⦿	✗	✗	✗	✗
Ours	✓	✓	✓ _{CR}	✗	✗

Table 1: Comparison with methodology-oriented causality surveys. ✓ indicates systematic coverage, ⦿ indicates coverage of only a few relevant studies, and ✗ indicates no systematic coverage. (Wang et al., 2024b) mainly covers counterfactual generation (CG), whereas ours covers broader counterfactual reasoning (CR).

needs, and organize it into semantic, intervention, and counterfactual causality. For each category, the taxonomy characterizes the core causal question, required model capabilities, representative tasks, and practical uses in causality analysis (§ 2). We then analyze existing studies, covering both evaluation-oriented and method-oriented research, and discuss their limitations (§ 3). Furthermore, unlike existing surveys that examine whether these benchmarks can truly assess LLMs’ contextual causality abilities or merely test their ability to retrieve knowledge (Yang et al., 2024b), we focus on the gap between them and real-world demands (§ 4). Lastly, we suggest promising directions, including real-world-oriented benchmark construction, more comprehensive and fine-grained evaluation, post-training data construction, and greater attention to intervention and counterfactual causality (§ 5). The comparison with existing surveys is summarized in Table 1. The resulting taxonomy of contextual causality is provided in Table 2.

2 Taxonomy of Contextual Causality

In this section, we introduce our taxonomy of contextual causality, which consists of three categories: semantic, intervention, and counterfactual causality. These categories can be viewed as a progression from causal understanding to causal prediction and further to alternative-world causal reasoning. Table 2 provides a detailed overview of the taxonomy. Prompt examples for representative tasks are provided in Appendix A.

Semantic Causality This category concerns causal relations that are already expressed or implied in a given natural language context, focusing on what can be seen from the given text (Drury

et al., 2022). In this form of causality, causal relations are conveyed through explicit lexical markers or implicit semantic relations in the text, and can be identified from the given context alone. To understand semantic causality, LLMs need three major capabilities, including causal identification, causal explanation, and causal strength estimation.

Causal identification refers to the ability of LLMs to understand the complex causal structures expressed or implied in text. Tasks under this capability mainly include ECI (Mirza et al., 2014), causal judgment (CJ) (Du et al., 2022), causal attribution (CA) (Ho et al., 2022), and causal detection and span extraction (CDS) (Ding et al., 2025). ECI aims to determine whether there is a causal relationship between two events, thereby supporting causal graph construction in practice. CJ focuses on selecting the most plausible cause or effect of a target event from candidate options, supporting causal link refinement. That is, when multiple possible causal links are available, LLMs must determine which link is most likely to represent a valid causal relation. CA seeks to identify which event or factor is responsible for an observed outcome, corresponding to responsibility assignment in practical causality analysis. CDS focuses on detecting whether a text expresses causal meaning and, when applicable, extracting the textual spans corresponding to the cause and effect. In practice, it is mainly used to filter out texts lacking causal information and to provide cause-and-effect spans as textual evidence for the detected causal relation.

Causal explanation and causal strength estimation further extend semantic causality beyond the identification of causal structures. Causal explanation requires LLMs to explain the mechanism,

Core Causal Question	Capability	Task	Practical Use
<i>Semantic Causality</i>			
What causal information is expressed or implied in context?	Identification	Event Causality Identification	Causal Graph Construction
		Causal Judgment	Causal Link Refinement
		Causal Attribution	Responsibility Assignment
		Causal Detection and Span Extraction	Causal Filtering and Evidence Grounding
	Explanation	Causal Explanation Generation	Causal Mechanism Explanation
	Strength Estimation	Causal Strength Assessment	Quantifying Causal Influence
<i>Intervention Causality</i>			
What outcome would occur under a given intervention in context?	Effect Estimation	Effect Prediction	Decision Support
<i>Counterfactual Causality</i>			
What would have happened if conditions had been different in context?	Imagination	Counterfactual Reasoning	Retrospective Analysis

Table 2: A taxonomy-driven overview of contextual causality with LLMs. The “Core Question” column specifies the causal question addressed by each category. The “Capability” column summarizes the causal ability required of LLMs for each category. The “Tasks” column lists representative tasks corresponding to each capability. The “Practical Use” column describes how each task can be employed in practical causality analysis.

process or logic through which a cause leads to its effect. The representative task under this capability is causal explanation generation (CEG) (Ho et al., 2022), which aims to explain why a causal relation holds. In practice, CEG is mainly used to explain causal mechanisms, making causal relations more interpretable and useful for human analysis. Causal strength estimation requires LLMs to assess the degree to which one event causally influences or contributes to another. The representative task for this capability is causal strength assessment (CSA) (Romanou et al., 2023), which goes beyond binary causal judgments by estimating the strength of the causal contribution between events. In practice, CSA is mainly used for quantifying causal influence, such as ranking causal factors or assigning strength scores to causal relations.

Intervention Causality This category concerns the causal relationship between an introduced action and the outcome it produces in a given context. Here, “intervention” is used in a textual sense and differs from the classical causal inference notion of manipulating the underlying data-generating process. Unlike semantic causality, which focuses on understanding causal information already expressed or implied in the context, intervention causality asks what outcome would occur when an intervention is introduced into the given context, corresponding to the question “What if I do...?” Understanding intervention causality requires LLMs

to perform effect estimation, i.e., estimating the likely outcomes of different actions in the given context. The representative task in this category is effect prediction (EP) (Li et al., 2026), which concerns inferring the effect that a given action would have in a given contextual setting. In practice, EP supports decision-making by helping compare possible actions based on their expected consequences.

Counterfactual Causality This category concerns whether and how the outcome would differ if a different action, event, or condition had occurred in a given context. Unlike intervention causality, counterfactual causality asks whether an observed outcome would have been different if a different action had been taken in the past, corresponding to the question “What if I had done...?” Understanding counterfactual causality requires LLMs to perform imagination, i.e., constructing an alternative scenario that differs from the observed facts and reasoning about how the outcome would change. The representative task under this category is counterfactual reasoning (CR) (Yu et al., 2023), which aims to infer what would have happened under such hypothetical changes. In practice, CR supports retrospective analysis, such as reviewing clinical decisions, analyzing system failures, and examining whether an outcome could have been different under alternative conditions. Since CG has already been systematically surveyed in (Wang et al., 2024b), we do not discuss it further here.

3 Contextual Causality with LLMs

In this section, we review existing work in this area. We organize them into evaluation-oriented studies, including performance evaluation and benchmark construction for evaluating LLMs, and method-oriented studies, including training-free and training-based approaches. When a category lacks relevant studies, we omit that type.

3.1 Performance Evaluation

Performance evaluation is essential for revealing the limitations of LLMs and identifying directions for future improvement.

Semantic Causality Gao et al. (Gao et al., 2023) conduct the first systematic evaluation of various GPT-series models (Brown et al., 2020; Achiam et al., 2023) on ECI, CJ, and CEG. Their results show that these models suffer from severe causal hallucination in both ECI and CJ, tending to over-attribute causal relations even when no true causality exists. One possible explanation is that natural language contains a bias, as causal relations are often expressed with explicit cue words such as “lead to” and “therefore”, while non-causal event pairs are rarely explicitly described. As a result, models trained on natural language may be more sensitive to causal relations. Nevertheless, GPT-series models remain relatively strong at explaining causal relations, as their generated explanations tend to be more complete and detailed. Zhao and Yan (Zhao and Yan, 2026) show that causal hallucination can be more severe in smaller LLMs ($\leq 1.5\text{B}$ parameters) on ECI. Their analysis of model responses suggests that these smaller models may mainly learn superficial answer templates. They further conduct a robustness analysis by introducing incorrect information as distracting evidence and find that model performance drops significantly, indicating sensitivity to misleading information. Meanwhile, Takayanagi et al. (Takayanagi et al., 2024) find that causal hallucination also exists in CDS.

Beyond causal hallucination, recent work (Zhao et al., 2026) shows that different LLMs may exhibit distinct causal behaviors in ECI. For example, Qwen-series models (Bai et al., 2023) exhibit causal neglect, where they often refuse to infer a causal relation by claiming insufficient information, even when the provided context contains sufficient causal evidence. This behavior may be related to post-training data, such as instruction tuning, which can lead to more conservative responses.

Moreover, Yu et al. (Yu et al., 2025) show that current LLMs, including DeepSeek, GPT, Claude, and others (Team et al., 2024; Grattafiori et al., 2024; Liu et al., 2024; Guo et al., 2025), still have limited ability to handle contextual causality when the causal structure is complex, and the task goes beyond binary causal judgment. Although these models achieve high accuracy on short-context CJ, where the context usually contains only a few sentences and the causal structure is relatively simple, their performance drops sharply on CSA in CRAB. This performance drop suggests that current LLMs are strong on simple, short-context binary causal judgment, but still limited in complex causal analysis, such as cases where an event has multiple causes with different causal strengths under long contexts.

Counterfactual Causality Kıcıman et al. (Kıcıman et al., 2023) evaluate GPT-series models on CR and find that they can achieve strong performance, outperforming prior methods. Yu et al. (Yu et al., 2025) further evaluate a broad range of LLMs on CR. Their results show that advanced LLMs already perform very well on CR. However, the CRASS benchmark (Frohberg and Binder, 2022) used in these evaluations has a relatively simple structure, in which each example typically presents a single premise and asks what would have happened if that premise had not occurred. Therefore, strong performance on CRASS does not imply that LLMs have robust counterfactual reasoning abilities applicable to complex scenarios, as such scenarios often involve long contexts, multiple interacting events, and even unobserved factors.

Limitations Existing performance evaluations of LLMs on contextual causality still provide an incomplete picture of models’ causal abilities. First, current evaluations are highly imbalanced across different categories of contextual causality, mainly because benchmarks for intervention and counterfactual causality remain relatively scarce. Second, many studies evaluate only GPT-series models or a small set of LLMs, making their conclusions difficult to generalize across model families, especially since different models may exhibit distinct failure modes such as causal hallucination and neglect. Moreover, failure-mode analysis remains limited, even though different models may exhibit distinct failures such as causal hallucination and neglect. Third, existing evaluations often rely on final-answer accuracy, partly because fine-grained evaluation of intermediate reasoning

processes is difficult to automate. This assessment makes it difficult to identify the true reasons behind LLMs’ failures in contextual causality. Fourth, robustness analysis remains insufficient, with relatively few studies examining model behavior under misleading information. Finally, many existing benchmarks formulate contextual causal tasks as multiple-choice questions, which may allow models to rely on superficial shortcuts rather than genuinely understanding the underlying causal relations. Finally, many benchmarks remain simplified, so strong performance on them does not necessarily indicate that models have a deep understanding of contextual causality. Future evaluations should include more complex situations, such as long causal chains and events caused by multiple factors with various causal strengths.

3.2 Benchmark Construction for LLMs

Benchmark construction is essential for evaluating the contextual causal understanding of LLMs.

Semantic Causality Since existing QA datasets mostly focus on fact-seeking questions such as “who”, “what”, “when”, and “where”, Ho et al. (Ho et al., 2022) propose WikiWhy to support why-question answering. WikiWhy mainly covers CA and CEG. Similarly, although many benchmarks have been developed to evaluate LLMs’ mathematical, coding, and general reasoning abilities, unified benchmarks for event relation extraction remain limited. To address this gap, Gong et al. (Gong et al., 2025) propose EventRelBench, a large-scale benchmark for evaluating LLMs on event relation extraction, where causal relation extraction, i.e., ECI, is included as one of the core tasks.

Motivated by the fact that human causal judgments are often graded rather than binary, Romanou et al. (Romanou et al., 2023) propose CRAB, which can be used to evaluate whether LLMs can estimate the strength of causal relations. In addition, to better evaluate whether LLMs can distinguish temporal relations from causal relations, Miliani et al. (Miliani et al., 2025) propose ExpliCa, a CJ benchmark. ExpliCITA (Bondielli et al., 2025) further extends ExpliCa to Italian. Considering data contamination, Chi et al. (Chi et al., 2024) propose CausalProbe-2024, a CJ benchmark.

However, the above benchmarks are mainly constructed from news, Wikipedia, or human-constructed world-knowledge questions, which may still leave a gap relative to emerging real-world applications such as medical diagnostic records

and legal case analysis. As LLM-based agents (Zhou et al., 2023; Yang et al., 2024a; Boiko et al., 2023) are increasingly used to assist humans in solving practical tasks, recent benchmarks have begun evaluating the causal attribution of LLMs across the execution trajectories of agentic systems. Zhang et al. (Zhang et al., 2025b) propose Who&When, which aims to identify which agent and which step cause task failures in LLM-based multi-agent systems. Similarly, Cemri et al. (Cemri et al., 2025) construct MAST-Data to analyze failure modes of these systems. However, these datasets still rely on human annotation, which is costly and time-consuming. To reduce this annotation burden, Zhang et al. (Zhang et al., 2025a) propose AgenTracer, which automatically constructs failure-attribution data via counterfactual replay and program-fault injection. Recently, MP-Bench (In et al., 2026) has also been proposed to rethink failure attribution in these systems. These benchmarks bring causal attribution closer to practical agent settings, but their data are still mainly generated from simple tasks, whereas real-world tasks, such as software iteration, involve multi-file changes and error-log diagnosis.

Intervention Causality Due to the lack of Portuguese causal benchmarks for evaluating LLMs, Lasheras and Pinheiro (Lasheras and Pinheiro, 2025) propose CaLQuest. The benchmark includes EP as a core component, but its examples are still mainly based on commonsense causal questions.

Counterfactual Causality Since counterfactual reasoning is important for LLMs, Yu et al. (Yu et al., 2023) propose IfQA, a benchmark for open-domain question answering under counterfactual presuppositions. The benchmark is still mainly constructed from Wikipedia-based knowledge.

Multi-level Contextual Causality Although benchmarks have been proposed for different categories of contextual causality, they are often designed for isolated tasks and evaluated under separate contexts. This makes it difficult to assess whether LLMs can perform different types of contextual causality within the same scenario. To address this gap, Li et al. (Li et al., 2026) propose METER, a multi-level contextual causality benchmark that evaluates ECI, EP, and CR under a unified context. However, METER is mainly built by reorganizing existing benchmarks, so its underlying data sources still largely come from Wikipedia and news rather than real-world scenarios.

Limitations Although recent studies have ex-

panded benchmarks for evaluating LLMs on contextual causality, most benchmarks are not constructed from practical application scenarios such as clinical diagnosis records or legal case documents. As a result, these benchmarks cannot fully evaluate LLMs’ ability to handle contextual causality in practical applications. Moreover, recent benchmarks based on the execution trajectories of agentic systems bring evaluation closer to practical scenarios, but their data are still mainly generated by simple tasks rather than long-horizon tasks that may take humans days or even months to complete, such as software iteration or scientific research. This may be because long-horizon benchmarks for agentic systems have recently emerged.

3.3 Training-free

Training-free methods are an important class of approaches for improving LLMs’ performance, as they do not require substantial resources (Wei et al., 2022; Wang et al., 2022b; Du et al., 2023).

Semantic Causality Since LLMs may lack sufficient causal knowledge to directly identify event causality, Cai et al. (Cai et al., 2025) propose Dr.ECI, a decomposed reasoning framework that incorporates causal structures into ECI. It decomposes ECI into multiple sub-tasks and uses multi-agent collaboration among a Causal Explorer, a Mediator Detector, and Direct and Indirect Reasoners. This framework enables LLMs to better identify both explicit and implicit causal relations. Similarly, Zou et al. (Zou et al., 2025) propose MR-Balance, which improves ECI through multi-agent debate with role assignment. Zeng et al. (Zeng et al., 2026) propose an ECI framework based on fuzzy aggregation of multisource evidence. However, these multi-agent approaches often incur substantial computational and time overhead. Zhao et al. (Zhao et al., 2026) propose a framework based on potential outcomes and actual causality theory that effectively mitigates causal bias in LLMs. However, these methods rely on the LLMs’ ability to solve the decomposed sub-tasks. Recent work (Zhao and Yan, 2026) suggests that such frameworks may not generalize well to smaller models, and they require multiple calls to LLMs, leading to higher computational costs.

Beyond multi-agent collaboration, several works enhance LLMs’ causal knowledge through external knowledge sources. Su et al. (Su et al., 2025) propose LKCER, which constructs a concept-level heterogeneous event graph and uses LLMs to en-

rich causal knowledge beyond the original event mentions. KnowQA (Wang et al., 2024c) reformulates document-level ECI as a binary question answering task. It leverages document-level event structures and designs single- and multi-turn causal QA strategies to identify causal relations. Wang et al. (Wang et al., 2024a) introduce a synthetic control method that uses retrieved similar event contexts to construct counterfactual controls for ECI. In addition, some studies also focus on other tasks. For example, Ashwani et al. (Ashwani et al., 2024) propose CARE-CA, a context-aware reasoning framework with counterfactual analysis. Chi et al. (Chi et al., 2024) propose G²-Reasoner, which incorporates general knowledge and goal-oriented prompts into LLMs’ causal reasoning process. Perak et al. (Perak et al., 2024) further explore whether retrieval-augmented generation (RAG) can help LLMs perform CJ more effectively. However, these methods depend on the coverage of external knowledge sources, which may limit their generalization to new scenarios, as each new scenario requires corresponding causal knowledge that is time-consuming to collect and difficult to make complete.

Counterfactual Causality Since pretraining corpora of Code-LLMs contain many conditional structures, such as *if*, which naturally encode causal structures, Liu et al. (Liu et al., 2023) explore whether Code-LLMs can exhibit stronger causal abilities than text-only LLMs. Their results show that Code-LLMs outperform text-only LLMs on CR. This suggests that exposure to training data with more explicit causal structures may help models acquire stronger causal reasoning abilities.

Limitations Multi-agent-based methods may only be effective for specific models and often require multiple calls to LLMs, leading to higher computational cost. Methods based on external knowledge sources also rely heavily on the coverage of the provided knowledge. To generalize to different real-world scenarios, they would require a nearly complete causal knowledge base for each scenario, whose completeness is difficult to guarantee. These requirements make existing methods costly to apply and difficult to generalize.

3.4 Training-based

Training remains the most effective way to inject knowledge into LLMs and improve their performance (Achiam et al., 2023; Guo et al., 2025).

Semantic Causality To mitigate causal halluci-

nation in smaller models on ECI, Zhao and Yan (Zhao and Yan, 2026) empirically analyze the criteria for high-quality CoT traces. Based on these criteria, they design a pipeline to generate high-quality CoT traces and use them to fine-tune smaller models. Their results show that this method can effectively mitigate causal hallucination. In addition, Zhang et al. (Zhang et al., 2025a) not only construct TracerTraj-2.5K to evaluate LLMs’ causal attribution ability, but also train LLMs based on this dataset. Their results show that the trained LLMs achieve stronger causal attribution performance in execution trajectories of agentic systems. However, these methods are still task-specific.

Limitations Current fine-tuning datasets for contextual causality remain limited, with only a few available for training LLMs on specific tasks. These datasets are insufficient for comprehensively improving LLMs’ contextual causal abilities. A unified dataset for improving LLMs on contextual causality is still lacking. One possible reason is that prior work has lacked a unified perspective on organizing contextual causality and has not sufficiently analyzed its role in addressing key challenges in current AI development, such as helping to build agents that can better solve real-world tasks through improved causal attribution. This may also limit exploration of reinforcement learning with verifiable rewards (RLVR), even though many contextual causality tasks are easily verifiable.

4 Benchmark Analysis

In this section, we analyze existing benchmarks. Representative benchmarks are summarized in Table 3. Existing discussions (Yang et al., 2024b) have examined whether these benchmarks can truly assess LLMs’ contextual causality abilities or merely test their ability to retrieve knowledge. Our analysis focuses on the gap between current benchmarks and real-world demands.

From Table 3, we can observe that they are mostly constructed from news, Wikipedia, stories, papers, social media, and human-written commonsense questions. However, strong performance on them may not indicate whether LLMs can perform contextual causal reasoning in real-world scenarios, where causal evidence is distributed across multiple lengthy, dynamic, and domain-specific sources, and causal analysis often requires reasoning beyond binary judgments over many-to-many causal relations with varying strengths. For example, in

Benchmark	Source	Focus
<i>Semantic Causality</i>		
<i>Event Causal Identification</i>		
MAVEN-ERE (Wang et al., 2022a)	Wiki	Intrinsic
MECI (Lai et al., 2022)	Wiki	Intrinsic
ESL (Caselli and Vossen, 2017)	News	Intrinsic
CTB (Mirza et al., 2014)	News	Intrinsic
BECauSE (Dunietz et al., 2017)	News	Intrinsic
CaTeRS (Mostafazadeh et al., 2016)	Stories	Intrinsic
METER (Li et al., 2026)	W&N	Intrinsic
EventRelBench (Gong et al., 2025)	W&N	Intrinsic
<i>Causal Strength Assessment</i>		
CRAB (Romanou et al., 2023)	News	Intrinsic
<i>Causal judgment</i>		
COPA (Roemmele et al., 2011)	Human	Intrinsic
XCOPA (Ponti et al., 2020)	Human	Intrinsic
e-CARE (Du et al., 2022)	Human	Intrinsic
ExpliCa (Miliani et al., 2025)	Human	Intrinsic
ExpliCITA (Bondielli et al., 2025)	Human	Intrinsic
MOCA (Nie et al., 2023)	Papers	Intrinsic
CausalProbe-2024 (Chi et al., 2024)	News	Intrinsic
<i>Causal Attribution</i>		
WikiWhy (Ho et al., 2022)	Wiki	Intrinsic
TellMeWhy (Lal et al., 2021)	Stories	Intrinsic
Who&When (Zhang et al., 2025b)	Agents	Intrinsic
TracerTraj-2.5K (Zhang et al., 2025a)	Agents	Intrinsic
MP-Bench (In et al., 2026)	Agents	Intrinsic
MAST-Data (Cemri et al., 2025)	Agents	Intrinsic
<i>Causal Explanation Generation</i>		
e-CARE (Du et al., 2022)	Human	Intrinsic
WikiWhy (Ho et al., 2022)	Wiki	Intrinsic
ART (Bhagavatula et al., 2019)	Stories	Intrinsic
<i>Causal Detection and Span Extraction</i>		
FinCausal2020 (Mariko et al., 2020)	News	Intrinsic
UniCausal (Tan et al., 2023)	News	Intrinsic
CausalTalk (Ding et al., 2025)	Reddit	Intrinsic
<i>Intervention Causality</i>		
<i>Effect Prediction</i>		
METER (Li et al., 2026)	W&N	Intrinsic
CaLQuest (Lasheras and Pinheiro, 2025)	Human	Intrinsic
<i>Counterfactual Causality</i>		
<i>Counterfactual Reasoning</i>		
CRASS (Frohberg and Binder, 2022)	Human	Intrinsic
IfQA (Yu et al., 2023)	Wiki	Intrinsic
TimeTravel (Qin et al., 2019)	Stories	Intrinsic
METER (Li et al., 2026)	W&N	Intrinsic

Table 3: Representative benchmarks for contextual causality. “Focus” indicates whether a benchmark measures causality ability or how much the causality ability of current LLMs improves decision-making for real-world needs. “Intrinsic” denotes the former, while “Extrinsic” denotes the latter. “Wiki”, “W&N”, “Human”, and “Agents” denote Wikipedia-derived data, mixed Wikipedia and news sources, human-constructed based on world or commonsense knowledge, and agent-system traces or execution logs, respectively.

construction, practitioners need to analyze design documents, construction logs, and progress updates to identify the causes of delays and cost overruns, assess their relative impacts, and determine the main contributing factors.

In addition, recent benchmarks derived from

agent-system traces or execution logs are closer to practical needs, as they can be used to evaluate LLMs’ performance in identifying failure causes in agentic systems. However, these benchmarks are still built on simple tasks. In real-world applications, many tasks are long-horizon, often requiring days or even months of iterative work by humans, such as software development, where progress depends on requirements analysis, code modification, debugging, testing, and repeated revision, or scientific research, where hypotheses, experiments, and analyses evolve over multiple stages.

Table 3 shows that existing benchmarks primarily evaluate specific contextual causal abilities of LLMs, such as causal judgment or causal attribution. In other words, most benchmarks adopt an intrinsic evaluation setting, where the goal is to test whether an LLM can identify or explain causal relations in a given context. However, there remains a lack of extrinsic evaluation measuring how much the contextual causal ability of current LLMs can improve the decision-making quality of AI systems or humans. However, there remains a lack of extrinsic evaluation measuring how much the contextual causal ability of current LLMs can improve the decision-making quality of AI systems or humans. This is important because the goal of evaluating and improving contextual causal reasoning is not merely to obtain higher performance on intrinsic benchmarks, but to enable AI systems or humans to make more reliable decisions.

5 Future Directions

Real-world-oriented Benchmark Construction

Future benchmarks for contextual causality should be designed to better align with real-world needs. First, future benchmarks should draw more from practical domains such as construction and healthcare, rather than relying on news and Wikipedia. In addition, future benchmarks could be constructed from agent-system execution trajectories in long-horizon tasks (Xiong et al., 2026; Thai et al., 2025; Chen et al., 2025), where even humans may need days or months of iterative work to complete them. However, constructing such benchmarks first requires well-defined long-horizon benchmarks for agents, so that meaningful execution trajectories can be collected for causal attribution evaluation. Second, future benchmarks should include extrinsic settings that assess whether LLMs can leverage contextual causality to improve the decision-

making of AI systems or humans, for example, by evaluating whether LLM-generated causal graphs improve the decision quality of agents.

More Comprehensive and Fine-grained Performance Evaluation Future evaluations should provide a more complete picture of LLMs’ contextual causal abilities. First, they should cover a broader range of tasks, especially intervention and counterfactual causality, which remain less explored due to limited benchmarks. Second, they should include more diverse model families, since different LLMs may exhibit different causal failure modes, such as causal hallucination or neglect. Third, future evaluations should move beyond final-answer accuracy toward fine-grained evaluation (Zheng et al., 2025; Li et al., 2025a; Chern et al., 2023; Min et al., 2023). Recent work on process reward models (PRMs) provides useful methodological guidance for evaluating intermediate reasoning steps through step-level supervision and learned verifiers or reward models (Zheng et al., 2025). Fine-grained evaluation methods such as FActScore and FacTool further illustrate how model outputs can be decomposed into smaller units and verified individually (Chern et al., 2023; Min et al., 2023). For contextual causality, merely judging the final answer is insufficient, as models may obtain correct answers through memorized knowledge or task-specific shortcuts rather than genuine understanding of contextual causality. Therefore, future work should develop automated methods for evaluating the intermediate processes underlying LLM responses.

Post-training Data Construction Since prior studies have suggested that LLMs’ causal hallucination and neglect may be related to limitations in training data, and existing data mainly focus on individual settings, such as ECI or CA in agentic systems, future work should construct more comprehensive post-training data for contextual causality (Xu et al., 2025a,b). A promising direction is to build unified datasets covering representative tasks across semantic, intervention, and counterfactual causality. For each sample in the dataset, the response should also include sufficiently detailed intermediate steps (Ye et al., 2025; Chen et al., 2026; Li et al., 2025d), rather than only the final answer, so that LLMs can learn the underlying causal reasoning process and comprehensively improve their contextual causal abilities. In addition, when constructing fine-tuning data, future work should also consider effective data mixture (Li et al., 2025c),

quality filtering (Zhang et al., 2026), and reducing the distribution gap between the CoT data and target models (Zhao and Yan, 2026).

Causal Hallucination Attribution Causal hallucination is an important manifestation of the limited contextual causal abilities of current LLMs, yet its causes remain insufficiently studied. Future work should systematically investigate causal hallucination, drawing on broader studies of LLM hallucination (Dziri et al., 2022; Ji et al., 2023; Kalai et al., 2025). Such attribution can help identify when and why LLMs produce incorrect causal relations and provide guidance for developing targeted mitigation methods.

Need for More Attention to Intervention and Counterfactual Causality Previous research has mainly focused on semantic causality, partly because the community has been concerned with improving LLMs’ language understanding, knowledge use, and basic reasoning abilities. However, as LLMs are increasingly used to autonomously solve real-world tasks (Wei et al., 2026; Gupta et al., 2025), decision-making becomes more central. Intervention and counterfactual causality are closely tied to this goal, because they help LLMs estimate the effects of possible actions and reflect on how past decisions could be improved. Therefore, future research should place more emphasis on intervention and counterfactual causality, including benchmark construction, performance evaluation, and method development.

6 Conclusion

In this survey, we present an overview of contextual causality with LLMs. We first propose a taxonomy of contextual causality motivated by the needs of causal analysis. This taxonomy organizes contextual causality into semantic, intervention, and counterfactual causality, and further characterizes each category by its core causal question, required capability, representative tasks, and practical role in causality analysis. We then summarize and analyze existing studies and benchmarks. Our analysis reveals several key limitations in current research, including limited, coarse-grained performance evaluation, insufficient post-training data, limited attention to interventions and counterfactual causality, and a gap between current benchmarks and real-world needs. Lastly, we discuss several promising directions for future research. We hope this survey can provide a roadmap for this field, highlight its

importance, and promote future research toward solving real-world problems.

Limitations

This survey focuses on contextual causality in text-only scenarios. However, the proposed taxonomy can also be extended to multimodal settings. For example, in an embodied scenario, foundation models need to identify causal relations from visual observations, language instructions, and action histories, predict the effects of possible interventions, reason about counterfactual outcomes under alternative actions, and finally make reliable decisions in the environment. This suggests that contextual causality is also important for multimodal foundation models, where contextual causality may involve images, videos, and sensor observations in addition to text. In future work, we plan to extend this survey from LLMs to multimodal foundation models, including vision-language models (VLMs), vision-language-action models (VLAs), and vision-language-navigation models (VLNs). In addition, our taxonomy is grounded in classical causal theory and primarily serves as a unified framework for analyzing contextual causality in LLMs rather than extending causal theory itself. Strengthening the theoretical foundation of the taxonomy remains an important direction for future work.

Acknowledgments

We gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grants ALLRP 585937-23 and RGPIN-2025-05097, and Mitacs under Grant IT35587.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Dushyant Singh Sengar, Mayank Jindal, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: Can large language models truly understand causality? In *Proceedings of the AAAI Symposium Series*, volume 4, pages 2–9.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei

- Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Adib Bazgir, Amir Habibdoust, Yuwen Zhang, and Xing Song. 2025. Causal mas: A survey of large language model architectures for discovery and effect estimation. *arXiv preprint arXiv:2509.00987*.
- Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.
- Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.
- Alessandro Bondielli, Martina Miliani, Luca Paglione, Serena Auriemma, Lucia C Passaro, and Alessandro Lenci. 2025. Llm struggle on explicit causality in italian. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 83–94.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Ruichu Cai, Shengyin Yu, Jiahao Zhang, Wei Chen, Boyan Xu, and Keli Zhang. 2025. Dr. eci: Infusing large language models with causal knowledge for decomposed reasoning in event causality identification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9346–9375.
- 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.
- Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, and 1 others. 2025. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*.
- Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2026. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Science China Information Sciences*, 69(6):161101.
- Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, and 1 others. 2025. Scienceagent-bench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pages 96934–96990.
- Qing Cheng, Zefan Zeng, Xingchen Hu, Yuehang Si, and Zhong Liu. 2025. A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *ACM Computing Surveys*, 58(3):1–37.
- I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, and 1 others. 2023. Factool: Factuality detection in generative ai—a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.
- 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.
- Shaobo Cui, Zhijing Jin, Bernhard Schölkopf, and Boi Faltings. 2024. The odyssey of commonsense causality: From foundational benchmarks to cutting-edge reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16722–16763.
- Shaobo Cui, Luca Mouchel, and Boi Faltings. 2025. Uncertainty in causality: A new frontier. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8022–8044.
- Xiaohan Ding, Kaike Ping, Buse Çarık, and Eugenia Rho. 2025. A multi-level benchmark for causal language understanding in social media discourse. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28764–28778.
- Brett Drury, Hugo Gonalo Oliveira, and Alneu de Andrade Lopes. 2022. A survey of the extraction and applications of causal relations. *Natural Language Engineering*, 28(3):361–400.
- Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.
- Jesse Dunietz, Lori Levin, and Jaime G Carbonell. 2017. The because corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th linguistic annotation workshop*, pages 95–104.
- Nouha Dziri, Sivan Milton, Mo Yu, Osmar R Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285.

- Jörg Froberg 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.
- Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is chatgpt a good causal reasoner? a comprehensive evaluation. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 11111–11126.
- Jie Gong, Biaoshuai Zheng, and Qiwang Hu. 2025. Eventrelbench: A comprehensive benchmark for evaluating event relation understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9084–9099.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Rushil Gupta, Jason Hartford, and Bang Liu. 2025. Llm for experiment design in scientific domains: Are we there yet? In *ICML 2025 Generative AI and Biology (GenBio) Workshop*.
- Joseph Y Halpern. 2016. *Actual causality*. MIT Press.
- Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. 2022. Wikiwhy: Answering and explaining cause-and-effect questions. *arXiv preprint arXiv:2210.12152*.
- Yeonjun In, Mehrab Tanjim, Jayakumar Subramanian, Sungchul Kim, Uttaran Bhattacharya, Wonjoong Kim, Sangwu Park, Somdeb Sarkhel, and Chanyoung Park. 2026. Rethinking failure attribution in multi-agent systems: A multi-perspective benchmark and evaluation. *arXiv preprint arXiv:2603.25001*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.
- Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.
- 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.
- Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. Tellmewhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610.
- Uriel Anderson Lasheras and Vl dia 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.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.
- Pengfeng Li, Chen Huang, Chaoqun Hao, Hongyao Chen, Xiao-Yong Wei, Wenqiang Lei, and See-Kiong Ng. 2026. Meter: Evaluating multi-level contextual causal reasoning in large language models. *arXiv preprint arXiv:2604.11502*.
- Xin Li, Zhuo Cai, Shoujin Wang, Kun Yu, and Fang Chen. 2025b. A survey on enhancing causal reasoning ability of large language models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 399–416. Springer.
- Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025c. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25366–25394.
- Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025d. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

- Xiao Liu, Da Yin, Chen Zhang, Yansong Feng, and Dongyan Zhao. 2023. The magic of if: Investigating causal reasoning abilities in large language models of code. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9009–9022.
- Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Hao-liang Wang, Tong Yu, and 1 others. 2025. Large language models and causal inference in collaboration: A comprehensive survey. *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7668–7684.
- Jing Ma. 2025. Causal inference with large language model: A survey. *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5886–5898.
- Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (fincausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32.
- 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.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.
- Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the tempeval-3 corpus. In *Proceedings of the EACL 2014 workshop on computational approaches to causality in language (CAtoCL)*, pages 10–19.
- Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the fourth workshop on events*, pages 51–61.
- Allen Nie, Yuhui Zhang, Atharva Shailesh Amdekar, Chris Piech, Tatsunori B Hashimoto, and Tobias Gerstenberg. 2023. Moca: Measuring human-language model alignment on causal and moral judgment tasks. *Advances in Neural Information Processing Systems*, 36:78360–78393.
- Benedikt Perak, Slobodan Beliga, and Ana Meštrović. 2024. Incorporating dialect understanding into llm using rag and prompt engineering techniques for causal commonsense reasoning. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 220–229.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.
- Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053.
- 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.
- Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.
- 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.
- 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.
- Takehiro Takayanagi, Masahiro Suzuki, Ryotaro Kobayashi, Hiroki Sakaji, and Kiyoshi Izumi. 2024. Is chatgpt the future of causal text mining? a comprehensive evaluation and analysis. In *2024 IEEE International Conference on Big Data (BigData)*, pages 6651–6660. IEEE.
- Fiona Anting Tan, Xinyu Zuo, and See-Kiong Ng. 2023. Unicausal: Unified benchmark and repository for causal text mining. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 248–262. Springer.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

- Minh VT Thai, Tue Le, Dung Nguyen Manh, Huy Phan Nhat, and Nghi DQ Bui. 2025. Swe-evo: Benchmarking coding agents in long-horizon software evolution scenarios. *arXiv preprint arXiv:2512.18470*.
- Guangya Wan, Yunsheng Lu, Yuqi Wu, Mengxuan Hu, and Sheng Li. 2024. Large language models for causal discovery: Current landscape and future directions. *arXiv preprint arXiv:2402.11068*.
- Haoyu Wang, Fengze Liu, Jiayao Zhang, Dan Roth, and Kyle Richardson. 2024a. Event causality identification with synthetic control. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1725–1737.
- Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, and 1 others. 2022a. 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.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024b. A survey on natural language counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4798–4818.
- 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.
- Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024c. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Tianxin Wei, Ting-Wei Li, Zhining Liu, Xuying Ning, Ze Yang, Jiaru Zou, Zhichen Zeng, Ruizhong Qiu, Xiao Lin, Dongqi Fu, and 1 others. 2026. Agentic reasoning for large language models. *arXiv preprint arXiv:2601.12538*.
- Xing Wu, Shaoqi Peng, Jingwen Li, Jian Zhang, Qun Sun, Weimin Li, Quan Qian, Yue Liu, and Yike Guo. 2024. Causal inference in the medical domain: a survey: X. wu et al. *Applied Intelligence*, 54(6):4911–4934.
- Lei Xiong, Kun Luo, Ziyi Xia, Wenbo Zhang, Jin-Ge Yao, Zheng Liu, Jingying Shao, Jianlyu Chen, Hongjin Qian, Xi Yang, and 1 others. 2026. Autotoresearchbench: Benchmarking ai agents on complex scientific literature discovery. *arXiv preprint arXiv:2604.25256*.
- Zhangchen Xu, Yang Liu, Yueqin Yin, Mingyuan Zhou, and Radha Poovendran. 2025a. Kodcode: A diverse, challenging, and verifiable synthetic dataset for coding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6980–7008.
- Zhangchen Xu, Adriana Meza Soria, Shawn Tan, Anurag Roy, Ashish Sunil Agrawal, Radha Poovendran, and Rameswar Panda. 2025b. Toucan: Synthesizing 1.5 m tool-agentic data from real-world mcp environments. *arXiv preprint arXiv:2510.01179*.
- 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.
- John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024a. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652.
- Linying Yang, Vik Shirvaikar, Oscar Clivio, and Fabian Falck. 2024b. A critical review of causal reasoning benchmarks for large language models. *arXiv preprint arXiv:2407.08029*.
- Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.
- Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. 2025. Causaleval: Towards better causal reasoning in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12512–12540.
- 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.
- Zefan Zeng, Qing Cheng, Xingchen Hu, Wentao Li, Weiping Ding, and Zhong Liu. 2026. Zero-shot event causality identification via multisource evidence fuzzy aggregation with large language models. *IEEE Transactions on Fuzzy Systems*.
- Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. 2025a. Agentracer: Who is inducing failure in the llm agentic systems? *arXiv preprint arXiv:2509.03312*.

Ruichen Zhang, Rana Muhammad Shahroz Khan, Zhen Tan, Dawei Li, Song Wang, and Tianlong Chen. 2026. The quest for efficient reasoning: A data-centric benchmark to cot distillation. In *International Conference on Learning Representations*, volume 2026, pages 94555–94578.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and 1 others. 2025b. Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems. *arXiv preprint arXiv:2505.00212*.

Yiheng Zhao, Yuanliang Li, Shreya Savant, and Jun Yan. 2026. Mitigating causal bias in llms via potential outcomes framework and actual causality theory. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4212–4222.

Yiheng Zhao and Jun Yan. 2026. Generating effective cot traces for mitigating causal hallucination. *arXiv preprint arXiv:2604.12748*.

Congming Zheng, Jiachen Zhu, Zhuoying Ou, Yuxiang Chen, Kangning Zhang, Rong Shan, Zeyu Zheng, Mengyue Yang, Jianghao Lin, Yong Yu, and 1 others. 2025. A survey of process reward models: From outcome signals to process supervisions for large language models. *arXiv preprint arXiv:2510.08049*.

Guanglin Zhou, Shaoan Xie, Guang-Yuan Hao, Shiming Chen, Biwei Huang, Xiwei Xu, Chen Wang, Liming Zhu, Lina Yao, and Kun Zhang. 2025. Emerging synergies in causality and deep generative models: A survey. *IEEE Transactions on Artificial Intelligence*.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, and 1 others. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Xiang Zou, Xuanhong Li, Po Hu, and Ming Dong. 2025. Mrbalance: A framework for enhancing event causality identification in multi-agent debates via role assignment. *Knowledge-Based Systems*, page 114470.

A Prompt Examples for Contextual Causality Tasks

Examples of prompts for representative contextual causality tasks are shown below. We use ESL (Caselli and Vossen, 2017) for ECI, CRAB (Romanou et al., 2023) for causal strength assessment, e-CARE (Du et al., 2022) for causal judgment and causal explanation generation, Who&When (Zhang et al., 2025b) for causal attribution, FinCausal2020 (Mariko et al., 2020) for causal detection and span extraction, and METER (Li et al., 2026) for effect prediction and counterfactual reasoning.

Task: Event Causal Identification

Prompt example from ESL (Caselli and Vossen, 2017):

Jenkin, who was “arrested” at Millom Pier at around 9.35am, is also facing a charge of “causing” unnecessary suffering to an animal in relation to the dog’s death.

Is there a causal relationship between “causing” and “arrested”?

Task: Causal Strength Assessment

Prompt example from CRAB (Romanou et al., 2023):

Article A: From Mexico in the north to Chile in the south, the region’s constantly see-sawing political map once again resembles that of the early 2000s, when a so-called “pink tide” of left-leaning governments washed over it. But analysts say this time is different: the trend is driven by pragmatism rather than ideology.

Article B: In the 2000s, a wave of left-wing electoral success swept Latin America, with a varied and complex turn to more progressive economic or social policies. In the 2010s, a right-wing political and conservative trend countered this tide, in clear reaction to the advances of progressives, and the right has dominated politics in the region for the past decade.

Event 1: A “pink tide” of left-leaning governments swept Latin America in the early 2000s. Event 2: In the 2010s, a right-wing political and conservative trend countered this tide.

What is the causality score between Event 1 and Event 2 from 0 to 100? Score above 80: Event 1 is definitely responsible for Event 2. Score between 50–80: Event 1 might have been responsible for Event 2. Score lower than 50: Events are somehow related but definitely not causally related.

Task: Causal Judgment

Prompt example from e-CARE (Du et al., 2022):

Input Event: The man fell unconscious. Question: Please select the cause of the input event from the following options. Option 1: The assailant struck the man in the head. Option 2: The assailant took the man’s wallet.

Which option is the cause of the input event?

Task: Causal Attribution

Prompt example from Who&When (Zhang et al., 2025b):

Context: The following is a failed multi-agent task-solving process. The system consists of three agents: Excel_Expert, BusinessLogic_Expert, and DataVerification_Expert.

Original task: This spreadsheet contains a list of clients for a retractable awning company. Each client has ordered a new awning for the back of their house within the last 90 days. The company makes different designs depending on whether the awning is made to block sunrises or sunsets. In this region, houses with odd-numbered street addresses face east, and houses with even-numbered street addresses face west. How many of these clients will be receiving the sunset awning design?

Agent system prompts:
Excel_Expert: Your role Excel_Expert: Expert in analyzing and processing data from Excel files. Task and skill instructions – Task Description: You will be responsible for scrutinizing, interpreting, and managing data from Excel files. The role requires a deep understanding of various Excel functions, including but not limited to pivot tables, data visualization, data cleaning, and advanced formulas. Skill Description: You possess exceptional skills in Excel, enabling you to manipulate,

analyze, and derive meaningful insights from complex datasets. You have the ability to automate repetitive tasks using VBA (Visual Basic for Applications) and generate comprehensive reports. Your expertise includes proficiency in handling large datasets, performing statistical analysis, and utilizing Excel's extensive array of tools to streamline data processing.

DataVerification_Expert: Your role DataVerification_Expert specializes in employing Bing Search API for powerful online search capabilities. With an adept skill set in navigating, parsing, and analyzing search results, this expert seamlessly extracts precise information from an array of text content. A critical part of their role involves vigilantly verifying the accuracy of this information, ensuring the use of reliable, trustworthy data.

Task and skill instructions – You will utilize the Bing Search API to execute specialized search queries, combining through the vast data available on the internet. Your skill in parsing and analyzing search results will be crucial in distilling the vast array of information to the most relevant details. When it comes to extracting specific information from text content, your ability to pinpoint and retrieve exact data points is paramount. An essential component of your work is verifying the accuracy of information, which requires a keen eye for detail and a robust methodology for cross-checking facts. With your combination of technical prowess using the Bing Search API and a meticulous approach to data verification, your role as the DataVerification_Expert is essential for any operation that demands high accuracy and reliability in information gathering. Useful instructions for task-solving – Follow the instructions provided by the user. Solve the task step by step if you need to. If a plan is not provided, explain your plan first. If the error cannot be fixed or the task is not solved even after the code executes successfully, analyze the problem, revisit your assumptions, collect any additional information you need, and consider a different approach. When you find an answer, verify the answer carefully. Include verifiable evidence in your response if possible. How to use code? Suggest Python code in a Python coding block or shell script in a sh coding block for the Computer_terminal to execute. When using code, you must indicate the script type in the code block. Do not suggest incomplete code that requires users to modify it. The last results will not be cached, so you need to provide all necessary information in a single code block. Do not use a code block if it is not intended to be executed by the Computer_terminal. The Computer_terminal cannot provide any other feedback or perform any other action beyond executing the code you suggest. The Computer_terminal cannot modify your code. Use the print function for the output when relevant. Check the execution result returned by the user. Do not ask users to copy and paste the result. If the result indicates an error, fix it and output the code again. If you want the Computer_terminal to save the code in a file before executing it, put #filename: <filename> inside the code block as the first line.

BusinessLogic_Expert: Your role BusinessLogic_Expert is an expert in understanding and applying business rules to data. This role involves understanding complex business requirements and translating them into actionable data logic aligned with organizational goals. **Task and skill instructions –** Your task involves analyzing business requirements and interpreting them into specific business rules that can be implemented within a data framework. You must possess a deep understanding of business processes and their relationship to data management and

manipulation. The ability to apply these rules accurately to data sets is crucial for ensuring the integrity and efficiency of business operations. Key skills include critical thinking, attention to detail, problem-solving, and proficiency with various data tools and platforms. You should be adept at both written and verbal communication to effectively convey business logic to stakeholders and team members. Additional knowledge in SQL, database design, and data analysis techniques will further enhance your capability in this role. By leveraging your expertise, you will contribute significantly to optimizing business operations through precise and effective data logic application.

Execution trace:

Step 0 – Excel_Expert: The Excel_Expert receives the general task and the manager's plan. It loads the Excel file from ../2023/validation/4d51c4bf-4b0e-4f3d-897b-3f6687a7d9f2.xlsx, identifies the Street Address column, extracts the first numeric substring from each address using str.extract with a digit-matching regular expression, converts it to an integer, filters even-numbered street addresses, and prints the number of such clients.

Step 1 – Computer_terminal: The code executes successfully. The terminal outputs the spreadsheet columns: ID, First Name, Last Name, Street Address, City, Zip, Phone, Email, and Carrier Pigeon. The terminal then outputs 4.

Step 2 – BusinessLogic_Expert: The BusinessLogic_Expert verifies that the column name is Street Address. It repeats the same reasoning: load the file, extract street numbers from Street Address, identify even-numbered addresses, and count the clients with even-numbered addresses. It provides essentially the same Python code and asks the terminal to run it again.

Step 3 – Computer_terminal: The code executes successfully again and outputs 4.

Step 4 – DataVerification_Expert: The DataVerification_Expert accepts the executed result. It states that the street-address column was correctly identified, the street numbers were accurately extracted, clients with even-numbered addresses were identified, and the final count of clients receiving the sunset awning design is 4.

Step 5 – DataVerification_Expert: TERMINATE.

The above multi-agent system produced an incorrect final answer. Which agent caused the task failure, and at which step did the decisive error occur?

Task: Causal Explanation Generation

Prompt example from e-CARE (Romanou et al., 2023):

Cause: The assailant struck the man in the head. **Effect:** The man fell unconscious.

Why can the cause lead to the effect?

Task: Causal Detection and Span Extraction

Prompt example from FinCausal2020 (Mariko et al., 2020):

Text: Things got worse when the Wall came down. GDP fell 20% between 1988 and 1993. There were suddenly hundreds of thousands of unemployed in a country that, under Communism, had had full employment.

Does the text express a causal relation? If yes, extract the cause span and the effect span.

Task: Effect Prediction

Prompt example from METER (Li et al., 2026):

On January 12, 2025, patient Robert Chen (Male, 58) was admitted to the Cardiac Care Unit at City General Hospital. Coronary angiography confirmed a throm-

botic occlusion in the proximal Left Anterior Descending (LAD) artery, causing the acute myocardial ischemia observed on the ECG. Regarding treatment options, clinical protocols define Percutaneous Coronary Intervention (PCI) as the standard procedure for mechanically recanalizing occluded vessels to re-establish blood flow. Furthermore, pathology reports indicate significant myocardial necrosis, which is time-dependent and developed progressively over the four-hour duration following his initial presentation at 09:00, during which the occlusion remained untreated.

Question: *If Percutaneous Coronary Intervention is performed, what is the expected outcome for the patient?*

Task: Counterfactual Reasoning

Prompt example from METER (Li et al., 2026):

On January 12, 2025, patient Robert Chen (Male, 58) was admitted to the Cardiac Care Unit at City General Hospital. Coronary angiography confirmed a thrombotic occlusion in the proximal Left Anterior Descending (LAD) artery, causing the acute myocardial ischemia observed on the ECG. Regarding treatment options, clinical protocols define Percutaneous Coronary Intervention (PCI) as the standard procedure for mechanically recanalizing occluded vessels to re-establish blood flow. Furthermore, pathology reports indicate significant myocardial necrosis, which is time-dependent and developed progressively over the four-hour duration following his initial presentation at 09:00, during which the occlusion remained untreated.

Would the extent of myocardial necrosis have been reduced if the occlusion had been detected during the initial screening?

B Literature Collection Scope

We collected literature according to the role of each study in this survey. For comparison surveys, we considered works from major NLP venues such as EMNLP and NAACL, broader AI and data mining venues such as KDD, AI-related journals such as IEEE Transactions on Artificial Intelligence and ACM Computing Surveys, as well as highly relevant preprints that have not yet been formally published. For method-oriented studies, we primarily considered papers from major NLP venues, including ACL, EMNLP, NAACL, EACL, and COLING, while also including highly relevant studies from broader AI venues. For datasets and benchmarks, we mainly included representative and widely used resources relevant to the contextual causality tasks discussed in this survey.