Causal Inference with Large Language Model: A Survey

Jing Ma
Department of Computer and Data Sciences, Case Western Reserve University
jing.ma5@case.edu

Abstract

Causal inference has been a pivotal challenge across diverse domains such as medicine and economics, demanding a complicated integration of human knowledge, mathematical reasoning, and data mining capabilities. Recent advancements in natural language processing (NLP), particularly with the advent of large language models (LLMs), have introduced promising opportunities for traditional causal inference tasks. This paper reviews recent progress in applying LLMs to causal inference, encompassing various tasks spanning different levels of causation. We summarize the main causal problems and approaches, and present a comparison of their evaluation results in different causal scenarios. Furthermore, we discuss key findings and outline directions for future research, underscoring the potential implications of integrating LLMs in advancing causal inference methodologies.

1 Introduction

1.1 NLP, LLM, and Causality

Causal inference is an important area to uncover and leverage the causal relationships behind observations, enabling a deep understanding of the underlying mechanism and potential interventions in real-world systems. Different from most classical statistical studies, causal inference presents unique challenges due to its focus on "causation instead of correlation", which intricates a complicated integration of human knowledge (e.g., domain expertise and common sense), mathematics, and data mining. Due to the inherent proximity to the human cognitive process, causal inference has become pivotal in many high-stakes domains such as healthcare (Glass et al., 2013), finance (Atanasov and Black, 2016), and science (Imbens and Rubin, 2015).

Traditional causal inference frameworks, such as structural causal model (SCM) (Pearl, 2009) and potential outcome framework (Imbens and Rubin, 2015) have systematically defined causal concepts, quantities, and measures, followed up with multiple data-driven methods to discover the underlying causal relationships (Spirtes and Zhang, 2016; Nogueira et al., 2022; Vowels et al., 2022) and estimate the significance of causal effects (Winship and Morgan, 1999; Yao et al., 2021). Despite their success, existing causal methods still fall short of matching human judgment in several key areas, such as domain knowledge, logical inference, and cultural context (Kıcıman et al., 2023; Zečević et al., 2023; Jin et al., 2023a). Besides, most traditional causal inference approaches only focus on tabular data, lacking the ability to address causality in natural language. However, the demand for causal inference in natural language has persisted, offering numerous potential applications. For example, clinical text data from electronic health records (EHR) holds valuable causal information for healthcare research. Causal inference in NLP is a promising direction, offering both challenges and benefits. Advancements in large language models (LLMs) provide new opportunities to enhance traditional methods, bridging the gap between human cognition and causal inference (Feder et al., 2022).

1.2 Challenges of Causal Inference in NLP

Although LLMs have shown eye-catching success, causal inference poses unique challenges for LLM capabilities. Unlike regular data, natural language is unstructured, high-dimensional, and large-scale, making traditional causal methods ineffective. Besides, causal relationships in the text are often obscure, and the complex semantics require advanced language models to uncover them. These challenges create significant hurdles for causal tasks in NLP and require new approaches, presenting transformative opportunities for advancing causal inference research.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 5901–5913

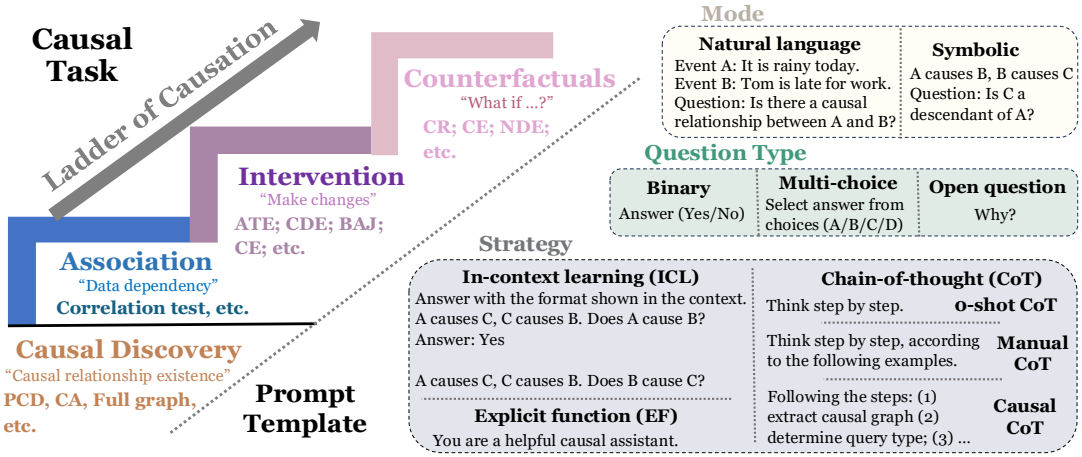

Figure 1: Representative causal tasks, their positions in the causal ladder, and examples of prompts w.r.t. mode, question type, and prompting strategy. PCD = pairwise causal discovery; CA=causal attribution; ATE=average treatment effect; CDE=controlled direct effect; BAJ=backdoor adjustment; CE=causal explanation; CR=counterfactual reasoning; NDE=natural direct effect.

1.3 Benefits Brought by LLMs

Despite the challenges, the increasing sophistication of LLMs has enhanced their ability in causal inference from linguistic data. LLMs bring the following key benefits to causal inference:

Domain knowledge. Traditional causal methods focus on numerical data, but domain knowledge is crucial in fields like medicine for identifying causal relationships. LLMs can extract this knowledge from large-scale text, reducing dependence on human experts for causal inference.

Common sense. LLMs can capture human common sense, which aids causal reasoning across contexts. For instance, legal cases require logic, and common sense often identifies abnormal events as causes (Kıcıman et al., 2023).

Sematic concept. Natural language, with its nuances and complexity, presents challenges for identifying causal relationships. Advances in NLP and LLMs, particularly in semantic modeling, offer new opportunities for deeper causal analysis.

Explainable causal inference. LLMs can provide tools for more intuitive, natural language-based explanations of causal reasoning, making complex concepts more accessible and enhancing user interaction with causal inference results.

1.4 Contribution and Uniqueness

Contribution. This survey systematically reviews existing research on using LLMs for causal inference with main contributions including: (1) We propose a clear categorization of studies, organized by tasks (Section 2) and technologies (Section 3). (2) We present a detailed comparison of existing LLMs (Section 4) and highlight key insights, connections, and observations. (3) We provide a comprehensive summary of benchmark datasets, focusing on key aspects for further study (Table 1). (4) We identify limitations and future research directions (Section 5), offering new perspectives on underexplored areas and opportunities.

Differences from existing surveys. Several previous surveys cover related topics (Liu et al., 2024; Kıcıman et al., 2023), while our survey differs as follows: (1) **Main scope:** We focus on "LLMs for causality," while other surveys with similar topic like (Liu et al., 2024), focus more on "causality for LLMs." (2) **Structure and content:** Our survey uniquely organizes research around tasks, methods, datasets, and evaluation, offering a clearer and more comprehensive review. (3) **Up-to-date:** We include the latest advancements, providing an up-to-date perspective on current trends and progress.

2 Preliminaries

2.1 Causality

Structural causal model. Structural causal model (SCM) (Pearl, 2009) is a widely used model to describe the causal relationships inside a system. An SCM is defined with a triple (U, V, F): U is a set of exogenous variables, whose causes are out of the system; V is a set of endogenous variables, which are determined by variables in $U \cup V$; $F = \{f_1(\cdot), f_2(\cdot), ..., f_{|V|}(\cdot)\}$ is a set of functions (a.k.a. *structural equations*). For each $V_i \in V$, $V_i = f_i(pa_i, U_i)$, where "$pa_i \subseteq V \setminus V_i$" and "$U_i \subseteq U$"

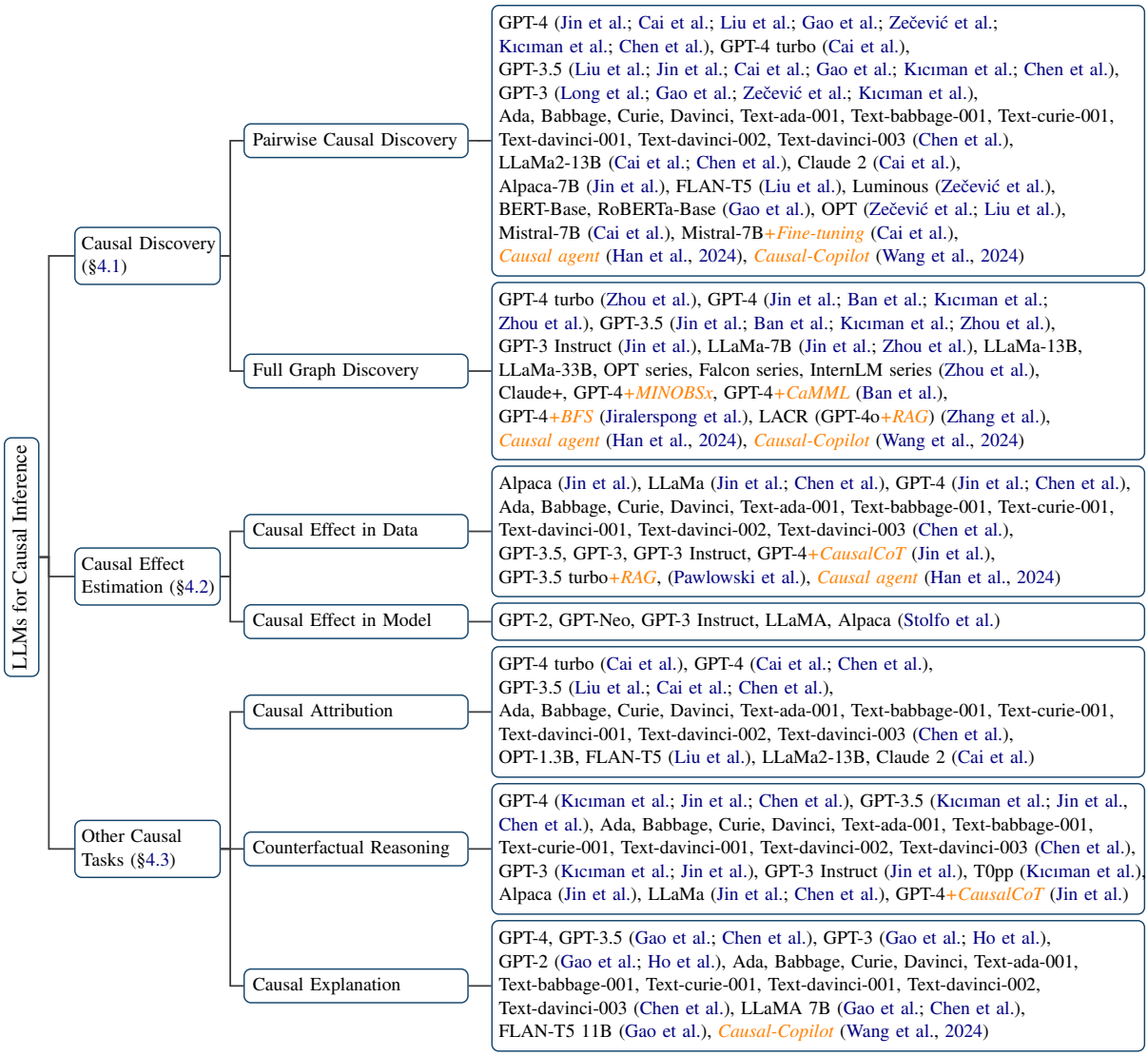

Figure 2: The major causal tasks and LLMs evaluated for these tasks. Noticeably, the citations in the figure correspond to the work of evaluations, rather than the original work of these models themselves. The strategies that are *not* merely based on prompting are highlighted in orange.

are variables that directly cause V_i. Each SCM is associated with a *causal graph*, which is a directed acyclic graph (DAG), where each node stands for a variable, and each arrow is a causal relationship.

Ladder of causation. The ladder of causation (Pearl and Mackenzie, 2018; Bareinboim et al., 2022) defines three rungs (Rung 1: *Association*; Rung 2: *Intervention*; Rung 3: *Counterfactuals*) to describe different levels of causation. Each higher rung indicates a more advanced level of causality. The first rung "Association" involves statistical dependencies, related to questions like "What is the correlation between taking a medicine and a disease?". Rung 2 "Intervention" moves further to allow interventions on variables, with questions like "If I take a certain medicine, will my disease be cured?". Rung 3 "Counterfactuals" relates to imagination or retrospection queries like "What if I had acted differently?", "Why?". Answering such questions requires knowledge related to the corresponding SCM. Counterfactual ranks the highest because it subsumes the first two rungs. A model that can handle counterfactual queries can also handle associational and interventional queries.

2.2 Causal Tasks and Related Rungs in Ladder of Causation

Causal inference involves various tasks. Figure 1 shows an overview of these tasks and their positions in the ladder of causation. We also show some examples of prompts w.r.t. mode, question type, and prompting strategy. We list several main causal tasks as follows:

Causal discovery. Causal discovery aims to infer

causal relationships from data. It includes discovering the causal graph and the structural equations associated with these causal relationships. Although causal discovery is not explicitly covered in the ladder of causation, it is often considered as "Rung 0" as it serves as a fundamental component in causal inference. Typical causal discovery questions include *pairwise causal discovery (PCD)* that only focuses on a pair of variables, and *full graph discovery* involving variables in the whole data system.

Causal effect estimation. Causal effect estimation (a.k.a. treatment effect estimation) aims to quantify the strength of the causal influence of a particular intervention or treatment on an outcome. In different scenarios, researchers may focus on the causal effect of different granularities, such as *individual treatment effect* (*ITE*, i.e., treatment effect on a specific individual), *conditional average treatment effect* (*CATE*, i.e., average treatment effect on a certain subgroup of population), *average treatment effect on the treated group* (*ATT*), and *average treatment effect* (*ATE*, i.e., average treatment effect on the entire population). Besides, people are also interested in the direct/indirect causal effects in certain scenarios, such as *natural direct effect* (*NDE*), *controlled direct effect* (*CDE*), and *natural indirect effect* (*NIE*). Another task related to causal effect estimation is *backdoor adjustment (BAJ)*, which aims to block all backdoor paths (Pearl, 2009) from the treatment to the outcome to exclude non-causal associations. Causal effect estimation tasks often span over Rung 2 and Rung 3.

Other tasks. There are many other tasks in causal inference. Among them, **causal attribution (CA)** refers to the process of attributing a certain outcome to certain events. **Counterfactual reasoning (CR)** investigates what might have happened if certain events or conditions had been different from what actually occurred. It explores hypothetical scenarios by considering alternative outcomes based on changes in "what if" circumstances. **Causal explanation (CE)** aims to generate explanations for an event, a prediction, or any causal reasoning process. This task often needs to answer causal questions in a specified human-understandable form or plain language. It is often in Rung 2 or 3, depending on the specific context. It is worth noting that, in many cases, different causal tasks may exhibit natural overlap in their scope. For instance, causal attribution and explanation commonly intersect with causal discovery and causal effect estimation. However, each task maintains a distinct focus.

3 Methodologies

Recent efforts (Kıcıman et al., 2023; Chen et al., 2024a; Gao et al., 2023) have explored leveraging LLMs for causal tasks. Unlike traditional data-driven or expert knowledge-based approaches, LLMs introduce novel methodologies, offering new perspectives for discovering and utilizing causal knowledge. Figure 2 lists LLMs developed or evaluated for causal tasks. We categorize current LLM methodologies for causal tasks as follows:

Prompting. Most existing works on causal reasoning with LLMs (Chen et al., 2024a; Kıcıman et al., 2023; Long et al., 2023; Jin et al., 2023a) focus on prompting, as it is the simplest approach. This includes regular strategies (like basic prompting, In-Context Learning (ICL) (Brown et al., 2020), and Chain-of-Thought (CoT) (Wei et al., 2022)) and causality-specific strategies (Jin et al., 2023a). For regular prompting, basic prompts (i.e., directly describing the question without any example or instruction) are most frequently used. There are also other efforts to devise more advanced prompting strategies. Among them, CaLM (Chen et al., 2024a) has tested 9 prompting strategies including basic prompt, adversarial prompt (Wallace et al., 2019; Perez and Ribeiro, 2022), ICL, 0-shot CoT (e.g., *"let's think step by step"* without any examples) (Kojima et al., 2022), manual CoT (i.e., guide models with manually designed examples), and explicit function (EF) (i.e., using encouraging language in prompts) (Chen et al., 2024a). Other works (Kıcıman et al., 2023; Long et al., 2023; Gao et al., 2023; Ban et al., 2023) also design different prompt templates. These works show substantial improvement potential of prompt engineering in causal reasoning tasks. For example, studies (Kıcıman et al., 2023; Chen et al., 2024a; Long et al., 2023) highlight that simple phrases like *"you are a helpful causal assistant"* can significantly boost performance. Additionally, there are causality-specific strategies, such as CausalCoT (Jin et al., 2023a), which combine CoT prompting with causal inference principles (Pearl and Mackenzie, 2018). Prompting-based methods can quickly and flexibly adapt to different causal tasks, but are still limited by the specificity of the prompt and thus easily lead to inconsistent causal responses.

Fine-tuning. Fine-tuning, as a widely recognized technique in general LLMs, is now also starting to gain attention for its application in causal tasks. Cai et al. (2023) propose a fine-tuned LLM for

Dataset	Task	Size (Unit)	Domain	Real	Sources	Citations
CEPairs (2016)	CD	108 (P)	Meteorology, etc.	R	37	(2016; 2023; 2023; 2023)
Sachs (2024)	CD	20 (R)	Biology	R	1	(2023; 2024; 2024)
Corr2Cause (2023b)	CD	200K (S)	Math	S	1	(2023b)
CLADDER (2023a)	Rung 1∼3	10K (S)	Dailylife, etc.	S	1	(2023a; 2023b)
BN Repo (2022)	CD	4∼84 (R)	Health, etc.	R	8	(2023)
COPA (2011)	CD	1K (Q)	Commonsense	R	1	(2023; 2011)
E-CARE (2022)	CD, CE	21K (Q)	Commonsense	R	1	(2023; 2022)
Asia (1988)	CD	8 (R)	Health	R	1	(1988; 2024; 2024)
CausalNet (2016)	CD	62M (R)	Web text	S	1	(2016; 2022)
CausalBank (2021)	CD	314M (P)	Web text	S	1	(2021; 2022)
WIKIWHY (2022)	CD,CE	9K (Q)	Wikipedia	R	1	(2022)
Neuro Pain (2019)	CD	770 (R)	Health	S	1	(2019; 2023; 2023)
Arctic Ice (2021)	CD	48 (R)	Climate	R	1	(2021; 2023)
CRASS (2022)	CR	275 (Q)	Commonsense	R	1	(2022)
CausalQA (2022)	CD, CE	1.1M (Q)	Web text, etc.	R	10	(2024; 2022; 2024)
CALM-Bench (2023)	CD, CA	281K (Q)	Science, etc.	R	6	(2023)
CausalBench (2024b)	Corr, CD	4∼195 (R)	Health, etc.	R	15	(2024b)
CaLM (2024a)	Rung 1∼3	126K (S)	Commonsense, etc.	R&S	20	(2024a)

Table 1: Datasets for LLM-related causal inference, with publication year, applicable tasks (CD=causal discovery; CR=counterfactual reasoning; CE=causal explanation), dataset size (as these datasets are not in a consistent form, we show the size w.r.t. different units, where P=causal pairs; R=causal relations; S=samples; Q=questions), domain, generation process (R: real-world; S: synthetic), number of data sources, and citations.

the pairwise causal discovery task (PCD, introduced in Section 4.1). This method generates a fine-tuning dataset with a Linear, Non-Gaussian, Acyclic Model (Shimizu et al., 2006), uses Mistral-7B-v0.2 (Jiang et al., 2023) as LLM backbone, and runs instruction finetuning with LoRA (Hu et al., 2021). The results achieve significant improvement compared with the backbone without fine-tuning. However, effective fine-tuning requires large computational resources, and may also suffer from overfitting problems.

Combining traditional causal methods. A line of studies combines LLMs with traditional causal methods. Considering causal inference often heavily relies on numerical reasoning, an exploration in Ban et al. (2023) leverages LLMs and data-driven causal algorithms such as MINOBSx (Li and Beek, 2018) and CaMML (O'Donnell et al., 2006). This method outperforms both original LLMs and data-driven methods, indicating a promising future for combining the language understanding capability of LLMs and the numerical reasoning skills of data-driven methods in complicated causal tasks. Jiralerspong et al. (Jiralerspong et al., 2024) combine LLM with a breadth-first search (BFS) approach for full graph discovery. It considers each causal relation query as a node expansion process, and gradually constructs the causal graph by traversing it with BFS. This method significantly reduces the time complexity from $O(n^2)$ to $O(n)$, where n is the number of variables. While it does not require access to observational data, their experiments show that the performance can be further enhanced with observational statistics. Recently, there have been increasing efforts on causality-driven LLM-based agents such as Causal Agent (Han et al., 2024) and Causal-Copilot (Wang et al., 2024) that automatically use LLMs to invoke causal tools for causal problems. This line of methods can potentially answer more complex causal questions than traditional methods and plain LLMs, but harmonizing between LLMs and data-driven causal approaches can also be a subtle problem.

Knowledge augmentation. LLMs with augmented knowledge can often better execute tasks for which they are not well-suited to perform by themselves, particularly for causal tasks that require domain-specific knowledge. Pawlowski et al. (2023) introduced two types of knowledge augmentation: context augmentation, which provides causal graphs and ITEs in the prompt, and tool augmentation, offering API access to expert systems for causal reasoning. Tool augmentation performs more robustly across varying problem sizes, as the LLM relies on the API for reasoning instead of reasoning through the graph itself. LACR (Zhang et al., 2024) applies retrieval augmented generation (RAG) to enhance the knowledge base of LLM for causal discovery, where the knowledge resources are from a large scientific corpus containing hidden insights about associational/causal relationships. Similar approaches (Samarajeewa et al., 2024) employ causal graphs as external sources for causal reasoning. Knowledge augmentation is

Model	CEPairs	E-CARE		COPA		CALM-CA	Neuro Pain
	Binary	Choice	Binary	Choice	Binary	Binary	Choice
ada	0.50	0.48	0.49	0.49	0.49	0.57	0.40
text-ada-001	0.49	0.49	0.33	0.50	0.35	0.48	0.50
Llama2 (7B)	-	0.53	0.50	0.41	0.35	0.32	-
Llama2 (13B)	-	0.52	0.50	0.44	0.36	0.42	-
Llama2 (70B)	-	0.52	0.44	0.50	0.45	0.49	-
Qwen (14B)	-	0.66	0.52	0.77	0.53	0.52	-
babbage	0.51	0.49	0.36	0.49	0.40	0.58	0.50
text-babbage-001	0.50	0.50	0.50	0.49	0.50	0.56	0.51
curie	0.51	0.50	0.50	0.50	0.50	0.58	0.50
text-curie-001	0.50	0.50	0.50	0.51	0.50	0.58	0.50
davinci	0.48	0.50	0.49	0.50	0.51	0.58	0.38
text-davinci-001	0.50	0.50	0.50	0.50	0.50	0.52	0.50
text-davinci-002	0.79	0.66	0.64	0.80	0.67	0.69	0.52
text-davinci-003	0.82	0.77	0.66	0.90	0.77	0.80	0.55
GPT-3.5-Turbo	0.81	**0.80**	0.66	**0.92**	0.66	0.72	0.71
GPT-4	-	0.74	**0.68**	0.90	**0.80**	**0.93**	**0.78**
0-shot ICL	-	0.83	0.71	0.97	0.78	0.90	-
1-shot ICL	-	0.81	0.70	0.93	0.76	0.90	-
3-shot ICL	-	0.71	0.70	0.80	0.81	0.91	-
0-shot CoT	-	0.77	0.68	0.91	0.79	0.92	-
Manual CoT	-	0.79	**0.73**	0.97	**0.82**	**0.95**	-
EF	-	**0.83**	0.71	**0.98**	0.80	0.92	**0.84**

Table 2: Performance (accuracy) of LLMs for causal discovery. Datasets include CausalEffectPairs (CEpairs), E-CARE, COPA, CALM-CA, and Neuro Pain. The white columns evaluate LLMs on PCD; in the gray column, on causal attribution; and in the cyan column, on full graph discovery. The upper part shows results with basic prompts; the lower part shows GPT-4 results with different prompting strategies. The tasks are shown in either binary "yes/no" or multi-choice formats. Results are drawn from Kıcıman et al. (2023) and Chen et al. (2024a).

especially effective for domain-specific tasks, as seen in RC^{2}R (Yu et al., 2024) for financial risk analysis and CausalKGPT (Zhou et al., 2024a) for aerospace defect analysis. However, high-quality causal knowledge sources are hard to collect, which can limit the LLM performance and also increase the method complexity.

4 Evaluations of LLMs in Causal Tasks

This section summarizes recent evaluation results of LLMs in causal tasks. We mainly focus on causal discovery and causal effect estimation, and also introduce several representative tasks spanning Rung 1∼3. A collection of datasets used in LLM-related causal tasks is shown in Table 1. In Table 2 and Table 4, we compare the performance of different LLMs in different tasks (including causal discovery and other tasks spanning different rungs) on multiple datasets. The mentioned LLMs include ada, babbage, curie, davinci (Brown et al., 2020), Qwen (14B) (Bai et al., 2023), text-ada-001, text-babbage-001, text-curie-001, text-davinci-001, text-davinci-002, text-davinci-003 (Ouyang et al., 2022), Llama 2 (7B, 13B, 70B) (Touvron et al., 2023), OpenAI's GPT series (Achiam et al., 2023; OpenAI, 2022), Mistral (7B) (Jiang et al., 2023), and Claude 2 (Models, 2023).

4.1 LLM for Causal Discovery

Unlike traditional causal discovery methods that rely solely on data values (Spirtes et al., 2000, 2013; Chickering, 2002), LLMs can also leverage metadata (e.g., variable names, problem context) to uncover implicit causal relationships, making their reasoning closer to human cognition. Many studies have explored LLMs for causal discovery (Kıcıman et al., 2023; Cai et al., 2023; Gao et al., 2023; Jin et al., 2023b; Long et al., 2023), focusing on pairwise causal discovery and full causal graph discovery, often framed as multi-choice or free-text

Model	Accuracy
GPT-3.5	0.58
GPT-4	0.76
GPT-4-Turbo	0.82
Llama2 (13B)	0.74
Claude 2	0.74
Mistral (7B) v0.2	0.75
Fine-tuned Mistral (7B) v0.2	**0.90**

Table 3: Accuracy of LLMs and a fine-tuned Mistral in PCD. The results are sourced from Cai et al. (2023).

question-answering tasks.

Pairwise causal discovery (PCD). PCD focuses on inferring the causal direction ($A \rightarrow B$ or $A \leftarrow B$) or determining the existence of a causal relationship. Kıcıman et al. (2023) use variable names in prompts, showing that LLMs (e.g., GPT-3.5, GPT-4) outperform state-of-the-art methods on datasets like CauseEffectPairs (Mooij et al., 2016) and domain-specific datasets like neuropathic pain (Tu et al., 2019). With proper techniques such as fine-tuning (as the example of fined-tuned Mistral shown in Table 3), the performance of PCD can be improved significantly. However, some studies (Zečević et al., 2023) suggest LLMs often act as "causal parrots", merely repeating embedded causal knowledge. Jin et al. (2023b) proposes a correlation-to-causation inference (Corr2Cause) task, where LLMs performed close to random. Although fine-tuning improved their performance, they still struggle with generalization in out-of-distribution scenarios. In summary, many studies Gao et al. (2023); Kıcıman et al. (2023); Jin et al. (2023b,a); Chen et al. (2024a) take a nuanced stance, acknowledging LLMs' strengths in PCD tasks while highlighting their limitations in reliably determining the existence of causal relationships.

Full causal graph discovery. Compared with PCD, identifying the full causal graph is a more complicated problem. In a preliminary exploration (Long et al., 2023), GPT-3 shows good performance in discovering the causal graph with 3-4 nodes for well-known causal relationships in the medical domain. In more complicated scenarios, the ability of different versions of GPT to discover causal edges (Kıcıman et al., 2023) has been validated on the neuropathic pain dataset (Tu et al., 2019) with 100 pairs of true/false causal relations. LLM-based discovery (GPT-3.5 and GPT-4) on Arctic sea ice dataset (Huang et al., 2021) has comparable or even better performance than representative baselines including NOTEARS (Zheng et al., 2018) and DAG-GNN (Yu et al., 2019). In Ban et al. (2023), the combination of the causal knowledge generated by LLMs and data-driven methods brings improvement in causal discovery in data from eight different domains with small causal graphs (5∼48 variables and 4∼84 causal relations). However, similar to PCD, LLMs also face many debates about their true ability to discover full graphs (Zhou et al., 2024b; Jin et al., 2023b).

4.2 LLM for Causal Effect Estimation

Although comparatively underexplored, LLMs have also shown impressive performance in causal effect estimation. These works can be mainly categorized into two branches: (1) **Causal effect in data:** LLMs estimate causal effects within data (Lin et al., 2023; Kıcıman et al., 2023) by leveraging their reasoning capabilities and large-scale training data. CLADDER (Jin et al., 2023a) benchmarks LLMs for causal effect estimation tasks (e.g., ATE in Rung 2, and ATT, NDE, NIE in Rung 3). Although this task remains challenging, techniques like CoT prompting (Jin et al., 2023a) significantly improve performance. (2) **Causal effect in models:** This branch investigates causal effects involving LLMs themselves, such as the impact of input data, neurons, or learning strategies on predictions (Vig et al., 2020; Meng et al., 2022; Stolfo et al., 2022). These studies help understand LLM behavior and support bias elimination (Vig et al., 2020), model editing (Meng et al., 2022), and robustness analysis (Stolfo et al., 2022). For example, Stolfo et al. (2022) explores the causal effect of input (e.g., problem description and math operators) on output solutions in LLM-based mathematical reasoning. In Vig et al. (2020), a causal mediation analysis for gender bias is conducted in language models.

4.3 LLM for Other Causal Tasks

Experiments (Chen et al., 2024a; Jin et al., 2023a; Kıcıman et al., 2023) have shown that there are various other causal tasks that LLMs can bring benefits to. **(1) Causal attribution**: LLMs show their capability in attribution tasks (Kıcıman et al., 2023; Cai et al., 2023) typically in the forms of "why" or "what is the cause" questions. Related tasks also include identifying necessary or sufficient causes (Liu et al., 2023; Kıcıman et al., 2023). By embedding human knowledge and cultural common sense, the results show that LLMs have the potential to

Model	CLADDER	CaLM			CLADDER	CaLM	CRASS	E-CARE
	Corr	ATE	CDE	BAJ	CR	NDE	CR	CE
ada	0.26	0.02	0.03	0.13	0.30	0.05	0.26	0.22
text-ada-001	0.25	0.01	0.01	0.29	0.28	0.01	0.24	0.33
Llama2 (7B)	0.50	0.03	0.02	0.18	0.51	0.03	0.11	0.42
Llama2 (13B)	0.50	0.01	0.01	0.19	0.52	0.02	0.20	0.39
Llama2 (70B)	0.51	0.09	0.09	0.13	0.52	0.13	0.17	0.42
Qwen (14B)	0.45	0.12	0.12	0.30	0.39	0.10	0.34	0.39
babbage	0.39	0.03	0.04	0.15	0.31	0.06	0.26	0.24
text-babbage-001	0.35	0.04	0.04	0.34	0.32	0.07	0.28	0.37
curie	0.50	0.01	0.04	0.23	0.49	0.01	0.22	0.30
text-curie-001	0.50	0.00	0.09	0.40	0.49	0.00	0.28	0.39
davinci	0.50	0.07	0.08	0.25	0.50	0.12	0.27	0.32
text-davinci-001	0.51	0.07	0.08	0.38	0.51	0.14	0.19	0.39
text-davinci-002	0.51	0.17	0.13	0.39	0.53	0.19	0.57	0.40
text-davinci-003	0.53	0.52	0.33	0.54	0.57	0.30	0.80	0.43
GPT-3.5-Turbo	0.51	0.38	**0.40**	0.44	0.58	0.30	0.73	**0.51**
GPT-4	**0.55**	**0.60**	0.31	**0.74**	**0.67**	**0.42**	**0.91**	0.46
0-shot ICL	0.60	0.19	0.25	0.72	0.65	0.27	0.85	0.48
1-shot ICL	0.66	0.24	0.30	0.70	0.71	0.38	0.78	0.41
3-shot ICL	0.61	0.70	0.70	**0.75**	0.69	0.29	0.70	0.40
0-shot CoT	0.57	0.57	0.28	0.73	0.66	0.43	**0.90**	0.53
Manual CoT	**0.66**	**0.93**	**0.91**	0.69	**0.77**	**0.80**	0.89	0.48
EF	0.60	-	-	0.72	0.70	-	0.87	**0.53**

Table 4: Performance (accuracy) of LLMs in causal tasks across the ladder of causation (Rung 1∼3) on datasets including CLADDER, CaLM, CRASS, and E-CARE. The gray column shows results for Rung 1 (corr=correlation), the white columns for Rung 2 (ATE=average treatment effect; CDE = controlled direct effect; BAJ= backdoor adjustment), and the cyan columns for Rung 3 (CR=counterfactual reasoning; NDE=natural direct effect; CE=causal explanation). The upper part shows results with basic prompts, while the lower part presents GPT-4 results with different prompting strategies. Data is sourced from Chen et al. (2024a) and Jin et al. (2023a).

flexibly address attribution problems in specific domains (such as law and medicine) where conventional methods may fall short (Kıcıman et al., 2023). **(2) Counterfactual reasoning**: Recent studies (Kıcıman et al., 2023; Jin et al., 2023a) explore different counterfactual reasoning scenarios, which are often in "what if" questions. While this task is one of the most challenging causal tasks, the improvement in LLMs compared to other methods is noteworthy. **(3) Causal explanation**: Many recent works explore causal explanations with LLMs (Bhattacharjee et al., 2023; Gat et al., 2023; Cai et al., 2023; Gao et al., 2023). Despite ongoing debates regarding LLM's actual ability for causal reasoning, most empirical studies positively indicate that LLMs serve as effective causal explainers (Gao et al., 2023). Such achievement is powered by LLMs' capability of analyzing language logic and answering questions with natural language.

4.4 Main Observations and Insights

From the evaluation above and results shown in Table 2 ∼ Table 4, we summarzie the main observations as follows: (1) **Model performance:** In general, many LLMs exhibit impressive performance in various causal tasks, especially in causal discovery, even with basic prompts. In some cases, their performance can be comparable to or even surpass human-level reasoning (Kıcıman et al., 2023). However, as the task difficulty increases from Rung 1∼3, their performance becomes less satisfactory in higher-level complicated causal reasoning tasks (Chen et al., 2024a). (2) **Enhancement through proper techniques:** The performance of LLMs can be significantly enhanced with effective prompting strategies (such as few-shot ICL and CoT) and other techniques like fine-tuning. These approaches enable models to be more causality-focused, with improved ability of

leveraging context and adaptively following correct steps in different causal reasoning tasks. Additionally, these models can provide valuable insights through causal explanations. (3) **General patterns:** While no definitive laws determine model performance universally, certain trends are still observable. For instance, scaling laws suggest that larger models generally perform better, although this is not always that straightforward. These trends provide valuable insights that can guide the future design and development of models. (4) **Variability in model effectiveness:** There is currently no universally superior LLM or strategy for causal tasks, as their effectiveness can vary significantly depending on the specific scenario. These observations highlight the need for more nuanced and adaptable approaches. (5) **Common issues:** Current LLMs still struggle with many issues in causal tasks. For example, the answers often lack robustness and are sensitive to changes in prompts (Kıcıman et al., 2023; Jin et al., 2023a). Furthermore, these models frequently default to memorizing and repeating information rather than actual causal reasoning (Zečević et al., 2023), which can limit their effectiveness in complex causal scenarios. Besides, LLMs often fail to generate self-consistent causal answers, i.e., the answers from LLMs often conflict with each other. Ongoing debates about whether LLM truly performs causal inference also compel more in-depth analysis.

5 Discussion and Future Prospects

In general, LLMs offer intriguing perspectives on causal inference, but current research also reveals many limitations, pointing to potential directions for future work that could advance the field (Zhang et al., 2023; Kıcıman et al., 2023).

Involving human knowledge: A more comprehensive integration of human knowledge into LLMs can improve causal reasoning, enabling analysis across both general and specialized fields like finance, health, and law (Chen et al., 2024b).

Improving data generation: Real-world data often lacks verified causal relations and counterfactuals. LLMs can generate diverse, realistic data with reliable causal relationships, enriching datasets and improving causal reasoning model training.

Addressing hallucinations: In causal reasoning, hallucinations are commonly generated and difficult to detect, leading to misleading causal conclusions. Reducing them is essential to improve the reliability of LLM in causal tasks.

Improving explanation and interactivity: Developing interpretable and instructable LLMs for causal reasoning is crucial. Techniques like fine-tuning, probing, prompt engineering, and optimizing reasoning chains can foster more collaborative and controllable causal inference.

Exploring multimodal causality: Real-world causal scenarios often involve multiple modalities. Recent studies have begun exploring causality across different modalities, such as images (Li et al., 2024) and videos (Lam et al., 2024). Future research could further investigate these multimodal approaches to enhance causal reasoning.

Developing a unified causal benchmark: There is currently a lack of unified and widely recognized benchmarks for evaluating causal performance for LLMs. Creating a comprehensive benchmark would facilitate LLM assessment.

Advancing causality-specialized models: Most current methods use original LLMs without sufficient focus on causality-centric model designs. There is a significant opportunity for further research and development in specialized causal LLMs to deepen their causal understanding and improve their effectiveness in causal inference.

6 Limitations

In this survey, it is important to acknowledge certain limitations that shape the scope and focus of our review. First, our analysis is primarily centered on the application of LLMs for causal inference tasks, thereby excluding exploration into how causality is utilized within LLMs themselves. This decision provides a targeted perspective on leveraging LLMs to enhance causal inference but does not delve into the internal mechanisms or implementations of causal reasoning within these models.

Second, while we comprehensively examine the technical aspects and methodological advancements in using LLMs for causal inference, we do not extensively discuss ethical considerations or potential societal impacts associated with these applications. Ethical dimensions, such as fairness, bias mitigation, and privacy concerns, are critical in the deployment of AI technologies, including LLMs, and warrant dedicated attention and scrutiny in future research and applications. Addressing these limitations ensures a nuanced understanding of the opportunities and challenges in harnessing LLMs for causal inference while also advocating for responsible and ethical AI development and deployment practices.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Vladimir A Atanasov and Bernard S Black. 2016. Shock-based causal inference in corporate finance and accounting research. *Critical Finance Review*, 5:207–304.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. 2022. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl*, pages 507–556.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2023. Llms as counterfactual explanation modules: Can chatgpt explain black-box text classifiers? *arXiv preprint arXiv:2309.13340*.

Lukas Blübaum and Stefan Heindorf. 2024. Causal question answering with reinforcement learning. In *Proceedings of the ACM on Web Conference 2024*, pages 2204–2215.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hengrui Cai, Shengjie Liu, and Rui Song. 2023. Is knowledge all large language models needed for causal reasoning? *arXiv preprint arXiv:2401.00139*.

Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. 2024a. Causal evaluation of language models. *arXiv preprint arXiv:2405.00622*.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024b. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*.

David Maxwell Chickering. 2002. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498.

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.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv:2205.05849*.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

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.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is chatgpt a good causal reasoner? a comprehensive evaluation. *arXiv preprint arXiv:2305.07375*.

Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2023. Faithful explanations of black-box nlp models using llm-generated counterfactuals. *arXiv preprint arXiv:2310.00603*.

Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. 2013. Causal inference in public health. *Annual review of public health*, 34(1):61–75.

Kairong Han, Kun Kuang, Ziyu Zhao, Junjian Ye, and Fei Wu. 2024. Causal agent based on large language model. *arXiv preprint arXiv:2408.06849*.

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*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yiyi Huang, Matthäus Kleindessner, Alexey Munishkin, Debvrat Varshney, Pei Guo, and Jianwu Wang. 2021. Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere. *Frontiers in big Data*, 4:642182.

Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023a. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023b. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. 2024. Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

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

Ting En Lam, Yuhan Chen, Elston Tan, Eric Peh, Ruirui Chen, Paritosh Parmar, and Basura Fernando. 2024. Causalchaos! dataset for comprehensive causal action question answering over longer causal chains grounded in dynamic visual scenes. *arXiv preprint arXiv:2404.01299*.

Steffen L Lauritzen and David J Spiegelhalter. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194.

Andrew Li and Peter Beek. 2018. Bayesian network structure learning with side constraints. In *International conference on probabilistic graphical models*, pages 225–236. PMLR.

Zhiyuan Li, Heng Wang, Dongnan Liu, Chaoyi Zhang, Ao Ma, Jieting Long, and Weidong Cai. 2024. Multimodal causal reasoning benchmark: Challenging vision large language models to infer causal links between siamese images. *arXiv preprint arXiv:2408.08105*.

Zhongyang Li, Xiao Ding, Ting Liu, J Edward Hu, and Benjamin Van Durme. 2021. Guided generation of cause and effect. *arXiv preprint arXiv:2107.09846*.

Victoria Lin, Louis-Philippe Morency, and Eli Ben-Michael. 2023. Text-transport: Toward learning causal effects of natural language. *arXiv preprint arXiv:2310.20697*.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, et al. 2024. Large language models and causal inference in collaboration: A comprehensive survey. *arXiv preprint arXiv:2403.09606*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Stephanie Long, Tibor Schuster, Alexandre Piché, ServiceNow Research, et al. 2023. Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*.

Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *Fifteenth international conference on the principles of knowledge representation and reasoning*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

C. Models. 2023. Model card and evaluations for claude models.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102.

Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. 2022. Methods and tools for causal discovery and causal inference. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 12(2):e1449.

OpenAI. 2022. Introducing chatgpt. Accessed: 2023-05-11.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rodney T O'Donnell, Ann E Nicholson, Bin Han, Kevin B Korb, M Jahangir Alam, and Lucas R Hope. 2006. Causal discovery with prior information. In *AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings 19*, pages 1162–1167. Springer.

Nick Pawlowski, James Vaughan, Joel Jennings, and Cheng Zhang. 2023. Answering causal questions with augmented llms. *Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML)*.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Chamod Samarajeewa, Daswin De Silva, Evgeny Osipov, Damminda Alahakoon, and Milos Manic. 2024. Causal reasoning in large language models using causal graph retrieval augmented generation. In *2024 16th International Conference on Human System Interaction (HSI)*, pages 1–6. IEEE.

Marco Scutari. 2022. Bayesian network repository.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. 2006. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search*. MIT press.

Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. Springer.

Peter L Spirtes, Christopher Meek, and Thomas S Richardson. 2013. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*.

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schölkopf, and Mrinmaya Sachan. 2022. A causal framework to quantify the robustness of mathematical reasoning with language models. *arXiv preprint arXiv:2210.12023*.

Zhengwei Tao, Zhi Jin, Yifan Zhang, Xiancai Chen, Xiaoying Bai, Yue Fang, Haiyan Zhao, Jia Li, and Chongyang Tao. 2024. A comprehensive evaluation on event reasoning of large language models. *arXiv preprint arXiv:2404.17513*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ruibo Tu, Chao Ma, and Cheng Zhang. 2023. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. *arXiv preprint arXiv:2301.13819*.

Ruibo Tu, Kun Zhang, Bo Bertilson, Hedvig Kjellstrom, and Cheng Zhang. 2019. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. *Advances in Neural Information Processing Systems*, 32.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. 2022. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.

Xinyue Wang, Kun Zhou, Wenyi Wu, Fang Nan, and Biwei Huang. 2024. Causal-copilot: An autonomous causal analysis agent.

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

Christopher Winship and Stephen L Morgan. 1999. The estimation of causal effects from observational data. *Annual review of sociology*, 25(1):659–706.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46.

Guanyuan Yu, Xv Wang, Qing Li, and Yu Zhao. 2024. Fusing llms and kgs for formal causal reasoning behind financial risk contagion. *arXiv preprint arXiv:2407.17190*.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

Cheng Zhang, Dominik Janzing, Mihaela van der Schaar, Francesco Locatello, and Peter Spirtes. 2023. Causality in the time of llms: Round table discussion results of clear 2023. *Proceedings of Machine Learning Research vol TBD*, 1:7.

Yuzhe Zhang, Yipeng Zhang, Yidong Gan, Lina Yao, and Chen Wang. 2024. Causal graph discovery with retrieval-augmented generation based large language models. *arXiv preprint arXiv:2402.15301*.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. 2024. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8.

Bin Zhou, Xinyu Li, Tianyuan Liu, Kaizhou Xu, Wei Liu, and Jinsong Bao. 2024a. Causalkgpt: industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing. *Advanced Engineering Informatics*, 59:102333.

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