Causal Inference in Large Language Model: A Survey

Anonymous ACL submission

Abstract

Causal inference has been a pivotal challenge across diverse domains such as medicine and economics, demanding a complicated integration of human knowledge, numerical reasoning, and data processing capabilities. Recent advancements in natural language processing (NLP), particularly with the advent of large language models (LLMs), have introduced transformative 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 their causal problems, methodologies, and present comparison of their evaluation results in different scenarios. Furthermore, we discuss key findings, emerging trends, 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 in mathematical reasoning to automate knowledge discovery. Different from most classical statistical and AI studies, causal inference focuses on the causal relationships between variables instead of merely statistical dependencies. Due to the inherent proximity to the human cognitive process, causal inference has become pivotal in scientific investigations, and also advocated its crucial application across various AI-related domains. For example, investigating the causal relations between a specific treatment (e.g., medication) and an outcome (e.g., the recovery from a disease) can provide more valuable insights for medical practices than simple correlation analysis. Traditional causal inference frameworks, such as Pearl's structural causal model (SCM) [39] and Rubin's potential outcome framework [20] have systematically defined causal concepts, quantities, and measures, followed up with multiple data-driven methods to discover the underlying causal relationships [45, 37, 52] and estimate the significance of causal effects [55, 56]. Despite their success, there is still a gap between existing causal frameworks and human's causal judgment [25, 58, 22], covering different aspects including lack of human domain knowledge, logic inference, and cultural background. The burgeoning field of NLP has recently shed light on its potential to improve traditional causal inference problems. Recently, researchers have delved into causal inference within NLP, offering fresh perspectives to bridge the gap between human cognition and methodologies for causal inference.

In fact, the motivation for causal inference in NLP has persisted over an extended period, offering a multitude of potential applications. For example, clinical text data in electronic health records (EHR) contains a large amount of underlying causal knowledge that can be utilized for healthcare-related research. However, most traditional causal inference approaches only focus on tabular data, lacking ability to discover and utilize the causality inside natural language. In general, causal inference in NLP is a promising research path with strong motivation, which offers a spectrum of challenges and benefits concurrently.

1.2 Challenges of Causal Inference in NLP

Although LLMs have shown eye-catching success in various tasks, causal inference still presents many distinctive challenges for LLM capabilities. Different from regular data types, the nature of natural language brings difficulties in causal processing and analysis. Text data is often unstructured, high-dimensional, and large-scale, in which context traditional causal inference methods are not applicable. Besides, causal relations inside text are often obscure and sparse. The complicated semantic meaning and ambiguity hidden in text data

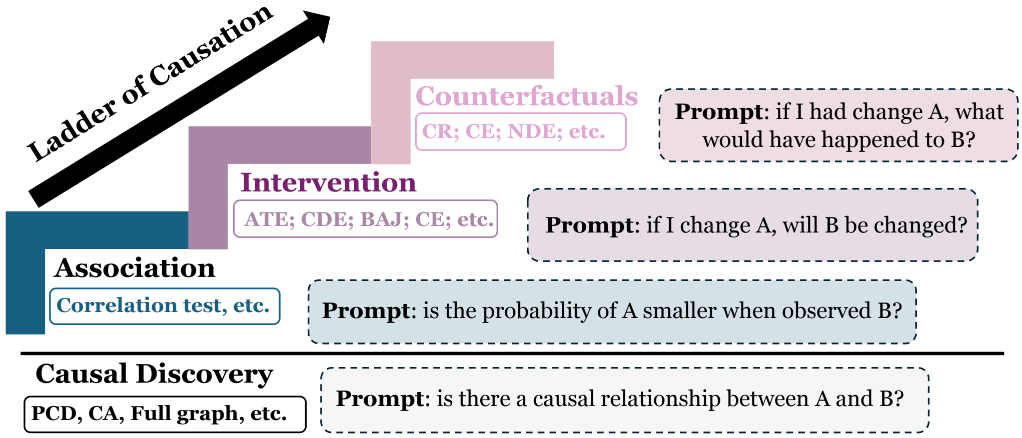

Figure 1: Representative causal tasks, their positions in the causal ladder, and examples of prompts. PCD = pairwise causal discovery; CA=causal attribution; ATE=average treatment effect; CDE=controlled direct effect; BAJ=backdoor adjustment; CE=causal explanation; CR=counterfactual reasoning; NDE=natural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

require sophisticated language modeling technologies to discover clear causal relationships, and also entail hurdles for other causal tasks such as causal intervention and counterfactual reasoning. These challenges demand new perspectives, assumptions, and technologies to address them effectively, offering revolutionary opportunities for current causal inference studies.

1.3 Opportunites that LLM Brings to Causal Inference

Despite the challenges, natural language has significant potential to yield advantages in causal inference. As NLP technologies and LLMs become increasingly sophisticated with diverse applications in recent years, the feasibility of understanding and unraveling causal relationships within linguistic data has been substantially improved. In general, LLM can bring benefits to causal inference in the following aspects:

Domain knowledge. Typical statistical methods for causal inference often only focus on the values of variables, while in many scenarios, domain knowledge plays an important role in causality-related tasks. More specifically, domain knowledge provides us with additional information to discover the true causal relationships and make meaningful interventions. For example, in many scientific domains such as medicine, incorporating the domain knowledge can draw conclusions that cannot be obtained solely through pure statistical methods, and expedite the development of relevant fields. However, collecting domain knowledge from human experts often demands considerable effort. Fortunately, the recent developments in NLP and LLM can extract domain knowledge from large-scale text information and thereby facilitate causal inference.

Common sense. Similar to domain knowledge, language models can serve as an effective tool to learn and utilize humans' general common sense to promote causal inference. As discussed in [25], a variety of common sense in different scenarios affects humans' recognition of causal relationships. For example, logical reasoning is essential for causal inference in law cases. Besides, abnormal events are often more likely to be recognized as causes for an outcome of interest in common sense.

Sematical concept. Compared with regular data types, natural language contains nuances, variations, and the richness of human expression, requiring advanced techniques for semantic analysis. Therefore, grasping clear causal concepts and relationships from text data is more challenging than other data types. Recent progress in NLP and LLM technologies, especially their ability in semantic modeling pave the way for in-depth causal studies in the next step.

Interactive and explainable causal inference. There have been long-lasting concerns about the difficult-to-understand terms and complicated reasoning processes in causal inference methods. LLMs such as ChatGPT have the potential to offer natural language-based interactive tools to promote human understanding for causal inference.

2 Preliminaries

2.1 Causality

Structural causal model. Structural causal model (SCM) [39] is a widely used model to describe the causal relationships inside a system. A 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 "$U_i \subseteq U$" are variables that directly cause V_i. Each SCM is associated with a causal graph, which is a directed acyclic graph (DAG). In the causal graph, each node stands for a variable, and each arrow represents a causal relationship.

Ladder of Causation. The ladder of causation [40, 3] 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 such as "What is the correlation between taking a medicine and a disease?". The second rung "Intervention" moves further to allow interventions on variables. Exemplar questions related to this rung are "What if I take a certain medicine, will my disease be cured?". The top rung "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 LLMs for causal inference tasks and their positions in the ladder of causation. We also show several examples of prompts corresponding to each rung. We list several main causal tasks which are most widely studied as follows:

Causal discovery. Causal discovery aims to infer causal relationships from data. It includes discovering a *causal graph* that describes the existence and direction of causal relationships inside a data system, as well as deriving the structural equations associated with these causal relationships. Although it is not officially covered in the ladder of causation, many works consider causal discovery as "Rung 0" as it serves as a fundamental component in causal inference.

Causal effect estimation. Causal effect estimation (a.k.a. treatment effect estimation) targets on quantifying the strength of the causal influence of a particular intervention or treatment on an outcome of interest. Causal effect estimation includes *experimental study* (where manipulation of variables is allowed) and *observational study* (without any manipulation). In different scenarios, researchers may focus on the causal effect of different granularities, ranging from *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), and *average treatment effect* (*ATE*, i.e., average treatment effect on the entire population). Causal effect estimation tasks often span over Rung 2 and Rung 3 in the ladder of causation.

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 human-understandable explanations for an event or a prediction, that is, answering the "why" questions in certain form or plain language. It is often in Rung 2 or Rung 3, depending on the specific context. In many cases, different causal tasks may exhibit natural overlap in their scope; for instance, attribution and explanation commonly intersect with causal discovery and causal effect estimation. However, each task maintains a distinct focus and emphasis.

3 Methodologies

Recently, there have emerged many efforts [25, 9, 15] to leverage LLMs for causal reasoning tasks. Different from traditional causal inference methodologies which are either data-driven or based on expert knowledge, the nature of LLM training and adoption introduces novel methodologies in causal inference, offering new perspectives and insights for discovering and utilizing causal knowledge in future research and applications. We summarize the current methodologies of LLM for causal tasks into the following categories:

Prompting. Most existing works [9, 25, 32, 22] of causal reasoning with LLMs focus on prompting, as it is the most straightforward method.

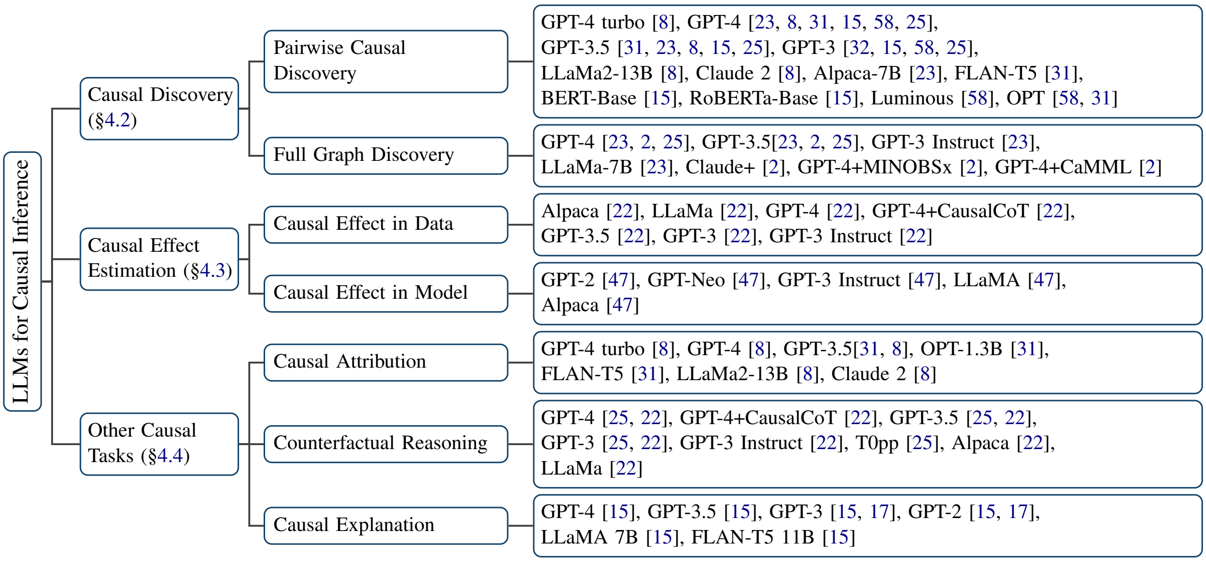

Figure 2: The major causal tasks and LLM models evaluated for these tasks. Noticeably, the citations in the figure correspond to the work with evaluations of different LLM models on specific tasks, rather than the original work of these models themselves.

Dataset	Year	Task	Size (Unit)	Domain	Real	Citations
CauseEffectPairs (37 datasets)[35]	2016	CD	108 (P)	Mixed	R	[35, 8, 58, 25]
Sachs [62]	2023	CD	20 (R)	Biology	R	[8, 62]
Corr2Cause [23]	2023	CD	200K (S)	Mixed	S	[23]
CLADDER [22]	2023	Effect, CR, CE	10K (S)	Mixed	S	[22, 23]
BN Repo [1]	2022	CD	4∼84 (R)	Mixed	R	[2]
COPA [42]	2011	CD	1,000 (Q)	Dailylife	R	[15, 42]
E-CARE [11]	2022	CD CE	21K (Q)	Mixed	R	[15, 11]
CausalQA [6]	2022	CD	1M (Q)	Mixed	R&S	[6]
CausalNet [33]	2016	CD	62M (R)	Mixed	S	[33, 11]
CausalBank [28]	2020	CD	314 M (P)	Mixed	S	[28, 11]
WIKIWHY [17]	2022	CD CE	9K (Q)	Mixed	R	[17]
Neuro Pain [50]	2019	CD	770 (R)	Health	S	[50, 25, 49]
Arctic Ice [19]	2021	CD	48 (R)	Climate	R	[19, 25]
CRASS [14]	2022	CR	275 (Q)	Mixed	R	[14]
CaLM [9]	2024	92 tasks in Rung 1∼3	126K (S)	Mixed	S	[9]

Table 1: Datasets for LLM-related causal inference, including publication year, applicable tasks (CD=causal discovery; Effect=causal effect estimation; CR=counterfactual reasoning; CE=causal explanation), dataset size (as different 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), and citations.

This line of work includes both regular prompting strategies (such as In-Context Learning (ICL) [7] and Chain-of-Thought (CoT) [54]) and causality-specific strategies. For regular prompting, most studies directly use a basic prompt (i.e., directly describe the question without any example or instruction). There are also other efforts to devise more advanced prompting strategies. Among them, CaLM [9] has tested 9 prompting strategies including basic prompt, adversarial prompt [53, 41], ICL, 0-shot CoT (e.g., "let's think step by step" without any examples) [26], manual CoT (i.e., guide models with manually designed examples), and explicit function (EF) (i.e., using encouraging language in prompts) [9]. Other works [25, 32, 15, 2] also design different prompt templates. These works show substantial improvement potential of prompt engineering in causal reasoning tasks. For example, results in [25, 9, 32] show adding simple sentences like "you are a helpful causal assistant" or "you are an expert in [DOMAIN NAME]" can impressively improve the causal inference performance for many models. Apart from these regular methods, other studies propose causality-specific prompting strategies. For example, CausalCoT [22] is a multi-step prompting strategy that combines CoT prompting and the causal inference engine [40].

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. [8] propose a fine-tuned LLM for the pairwise causal discovery task (PCD) (introduced in Section 4.2. This method generates a fine-tuning dataset with a Linear, Non-Gaussian, Acyclic Model [43], uses Mistral-7B-v0.2 [21] as LLM backbone, and runs instruction finetuning with LoRA [18]. The results achieve significant improvement compared with the backbone without fine-tuning.

Combining LLMs with data-driven causal methods. Considering causal inference tasks often heavily rely on numerical reasoning from data, another line of works combine LLMs with traditional data-driven causal methods. An exploration in [2] leverages LLMs and data-driven causal algorithms such as MINOBSx [27] and CaMML [38]. 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.

4 Evaluations of Causal Inference in LLM

4.1 Overview

In this section, we summarize recent progress in LLMs in causal tasks. We mainly focus on causal discovery and causal effect estimation, and also introduce several representative tasks spaning Rung 1 to Rung 3 in the ladder of causation. A collection of datasets used in LLM-related causal tasks is shown in Table 1. We also list the LLMs evaluated in the mentioned tasks and their corresponding evaluation papers in Figure 2.

4.2 LLM for Causal Discovery

Causal discovery aims to identify the causal relationships between different variables, which often serves as a fundamental step in real-world data analysis. Most traditional causal discovery approaches rely on the data values and use statistical approaches to infer the underlying causal structure over the corresponding variables. These approaches include *constraint-based methods* (e.g., PC algorithm [44] and FCI algorithm [46, 60]) which infer causal relationships by leveraging conditional independence tests, and *score-based methods* which assign scores to candidate causal graphs w.r.t. certain scoring criterion and seek the candidate causal graph with the highest score (e.g., GES algorithm [10]). Various classical statistical approaches and recent machine learning or deep learning technologies [45, 37, 52] have been used in causal discovery.

Recent developments in LLMs provide new perspectives for causal discovery [48, 30, 25]. Different from most existing causal discovery methods which can only utilize the data values of variables, LLMs can also leverage the metadata (e.g., the names of variables, the problem context) related to these variables to discover the implicit causal relationships. This reasoning process makes LLM-based causal discovery closer to human recognition. Recent literature [25] refer to this ability as knowledge-based causal discovery, and their experiments show that LLM-based knowledge-based causal discovery outperforms existing causal discovery methods on benchmarks [35]. Currently, a variety of investigations have been conducted on LLMs in causal discovery tasks [25, 8, 15, 23, 32]. These investigations are often conducted in the form of multi-choice or free-text question-answering, and they can mainly be divided into two types: pairwise causal discovery and full causal graph discovery.

Pairwise causal discovery ((PCD)) focuses on a pair of variables, either aiming to infer the causal direction ($A \rightarrow B$ or $A \leftarrow B$) between a given pair of variables (A, B), or aiming to judge the existence of a causal relation between two variables. Among them, the experiments in [25] use the names of variables when constructing prompts, and their results show that LLMs (including GPT-3.5 and GPT-4) outperform state-of-the-art methods both on datasets with common variables (e.g., CauseEffectPairs [35]) and datasets that require

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	-
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**
GPT-4 (0-shot ICL)	-	0.83	0.71	0.97	0.78	0.90	-
GPT-4 (1-shot ICL)	-	0.81	0.70	0.93	0.76	0.90	-
GPT-4 (3-shot ICL)	-	0.71	0.70	0.80	0.81	0.91	-
GPT-4 (0-shot CoT)	-	0.77	0.68	0.91	0.79	0.92	-
GPT-4 (Manual CoT)	-	0.79	**0.73**	0.97	**0.82**	**0.95**	-
GPT-4 (EF)	-	**0.83**	0.71	**0.98**	0.80	0.92	**0.84**

Table 2: Performance (accuracy) of different models in causal discovery tasks on different datasets, including CausalEffectPairs (CEpairs for short), E-CARE, COPA, CALM-CA, and Neuro Pain. In the columns in white (CausalEffectPairs, E-CARE, COPA), the models are evaluated for the pairwise causal discovery task; In the column in gray, the models are evaluated for the causal attribution task; in the column in cyan, the models are evaluated for the full graph discovery task. In the upper part, we show results with basic prompt; while in the lower part, we show results of GPT-4 with different prompting strategies. We also present results under prompts in the form of binary "yes/no" questions and multi-choice questions. The results are collected from Kıcıman et al. [25] and Chen et al. [9]. Note that the experimental settings such as prompt templates may be different.

particular domain knowledge (e.g., neuropathic pain [50]). Despite the encouraging results, the empirical analysis from [58] implies that in many cases, LLMs are just "causal parrots" that repeat the embedded causal knowledge. A comparison between ChatGPT and fine-tuned small pre-trained language models [15] shows LLMs' advantage in some causal discovery tasks, but this work also discusses that the ability of LLMs in determining the existence of a causal relationship is worse than simply selecting the cause or effect of an input event from given options. Jin et al. [23] proposes a correlation-to-causation inference (Corr2Cause) task to evaluate the causal inference performance of LLMs. Their experimental results reveal that LLM models perform almost close to random on the task, even though this issue could be mitigated through fine-tuning, these models still have limitations in generalization on out-of-distribution settings.

Full causal graph discovery aims to identify the full causal graph that describes the causal relationships among a given set of variables. Compared with pairwise causal discovery, discovering the full causal graph is a more complicated problem as it involves more variables. In a preliminary exploration [32], 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 [25] has been validated on the neuropathic pain dataset [50] with 100 pairs of true/false causal relations. LLM-based discovery (GPT-3.5 and GPT-4) on Arctic sea ice dataset [19] has comparable or even bet-

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
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
GPT-4 (0-shot ICL)	0.60	0.19	0.25	0.72	0.65	0.27	0.85	0.48
GPT-4 (1-shot ICL)	0.66	0.24	0.30	0.70	0.71	0.38	0.78	0.41
GPT-4 (3-shot ICL)	0.61	0.70	0.70	**0.75**	0.69	0.29	0.70	0.40
GPT-4 (0-shot CoT)	0.57	0.57	0.28	0.73	0.66	0.43	**0.90**	0.53
GPT-4 (Manual CoT)	**0.66**	**0.93**	**0.91**	0.69	**0.77**	**0.80**	0.89	0.48
GPT-4 (EF)	0.60	-	-	0.72	0.70	-	0.87	**0.53**

Table 3: Performance (accuracy) of different models in causal tasks in the ladder of causation (Rung 1 ∼ Rung 3) on different datasets, including CLADDER, CaLM, CRASS, and E-CARE. The column in gray correspond to tasks in Rung 1 (corr=correlation), the columns in white involve tasks in Rung 2 (ATE=average treatment effect; CDE = controlled direct effect; BAJ= backdoor adjustment); the columns in cyan correspond to tasks in Rung 3 (CR=counterfactual reasoning; NDE=natural direct effect; CE=causal explanation). In the upper part, we show results with the basic prompt; while in the lower part, we show results of GPT-4 with different prompting strategies. The results are collected from Chen et al. [9] and Jin et al. [22]. Note that the experimental settings such as prompt templates may be different.

ter performance than representative baselines including NOTEARS [61] and DAG-GNN [57]. In [2], 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). But similarly to pairwise causal discovery, LLMs also face many doubts and debates about their true ability in full causal graph discovery.

4.3 LLM for Causal Effect Estimation

Causal effect estimation is a task to quantify how much manipulating a treatment can causally influence an outcome. In most cases, the causal effect of interest is estimated from observational data. Researchers in the NLP community have also made lots of efforts in causal effect estimation from text data [13, 24]. Causal effect estimation on text data faces unique challenges due to the high-dimensional and complicated nature, for example, some important assumptions (e.g., positivity assumption [12]) in traditional causal effect estimation are easily violated when high-dimensional text information is a confounder [24]. Fortunately, the NLP progress in recent decades, such as word embeddings [1], topic modeling [5] and dependency parsing [36] have significantly contributed to estimating causal effects on text.

LLM has recently offered opportunities in causal effect estimation as well. Recently, the connection between causal effect estimation and LLMs includes two different branches: **(1) Causal Effect in Data**: In this task, LLMs aim to estimate the causal effect inside data [29, 25] by leveraging their reasoning capability and properties (e.g., ability in handling large-scale training corpora). A benchmark for the capability of LLMs in causal inference, CLADDER [22], includes query types regarding causal effect estimation at different levels, e.g., average treatment effect (ATE), average treatment effect on the treated (ATT), natural direct effect (NDE), and natural indirect effect (NIE). These queries cover the **Rung 2 (e.g., ATE)** and **Rung 3 (e.g., ATT, NDE, NIE)** of the Ladder of Causation [40, 3]. Existing evaluations show that the causal effect estimation task is still quite challenging for most LLMs. But an encouraging finding is that proper techniques such as chain-of-thought (CoT) prompting strategy [22] can improve the performance significantly. **(2) Causal Effect in Model**: This task aims to analyze the causal effect that involves the LLM model itself. Most commonly, we focus on the causal effect of input data, model neurons, or learning strategies on LLMs' predictions [51, 34, 47]. These studies can reveal the underlying LLM model behavior and promote further investigations such as bias elimination [51], model editing [34], and robustness quantification [47]. For example, [47] explores the causal effect of input (e.g., problem description and math operators) on output solutions in LLM-based mathematical reasoning. In [51], gender bias effects propagated from model input to output are detected and analyzed in language models.

4.4 LLM for Other Causal Tasks

There are various other causal inference tasks that LLMs can bring benefits to. **(1) Causal Attribution**: LLMs show their capability in attribution tasks [25, 8], which are often in the forms of "why" or "what is the cause" questions. Related tasks also include identifying necessary or sufficient causes [31, 25]. By embedding human knowledge and cultural common sense, LLMs have the potential to flexibly address attribution problems in specific domains (such as law, economics, and medicine) where conventional methods may fall short [25]. **(2) Counterfactual Reasoning**: Recent studies [25, 22] conduct experiments on LLMs in different counterfactual reasoning scenarios, which are often in "what if" questions. While this task is one of the most challenging tasks in causal inference, the demonstrated improvement in LLMs compared to other methods is noteworthy. **(3) Causal Explanation**: Many recent works investigate causal explanations based on queries on LLMs [4, 16, 8, 15]. Despite ongoing debates regarding LLM's actual ability for causal reasoning, most empirical studies positively indicate that LLMs serve as effective causal explainers [15]. Such achievement is powered by LLMs' capability of analyzing language logic and responding to questions using natural language.

In Table 2 and Table 3, we compare the performance of different LLMs in different tasks (including causal discovery and other tasks spaning Rung 1 to Rung 3) on multiple datasets. From the results, we notice that: many LLMs can achieve human-comparable performance in causal discovery even with basic prompts. Furthermore, with proper prompting strategies, the performance can be remarkably improved.

5 Discussion and Future Prospects

In general, as aforementioned, LLMs bring promising perspectives to causal inference, but there are also many limitations of current research and thus leave research directions in the future. First, a lot of literature [25, 22] have shown that the causal inference capability of LLMs is quite sensitive to the specific choice of prompts. Modifications in a few words and sentences can lead to significant changes in performance. Besides, LLMs often fail to generate self-consistent answers for causal queries, i.e., the answers from LLMs often present causal relationships that conflict with each other. Ongoing debates and criticisms about whether LLM truly performs causal inference also compel more in-depth and precise analysis and evaluation. Overall, there are many promising possibilities for the future of this research area [59, 25], including: (1) Incorporating domain knowledge into LLMs more comprehensively and intelligently, which holds the potential for interdisciplinary knowledge integration, discovery, and validation in specialized fields; (2) Natural language-based causal data generation, which augments the natural language in a causality-consistent manner to provide LLMs with more diverse and realistic data sources (3) Hallucination elimination in causal reasoning, ensuring more accurate and reliable causal inference; and (4) Interpretable and instructable causal reasoning, designing strategies for LLMs to interact with humans, providing the reasoning chains of LLMs and accept human instructions or feedback during the causal reasoning process, and fostering collaborative causal inference between humans and AI.

6 Limitations

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

Secondly, 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

[1] Felipe Almeida and Geraldo Xexéo. 2019. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.

[2] 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*.

[3] 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.

[4] 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*.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[6] 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.

[7] 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.

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

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

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

[11] 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*.

[12] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. 2021. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.

[13] 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.

[14] 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.

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

[16] 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*.

[17] 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*.

[18] 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*.

[19] 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.

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

[21] 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*.

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

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

[24] Katherine A Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. *arXiv preprint arXiv:2005.00649*.

[25] 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*.

[26] 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.

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

[28] 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*.

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

[30] Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. 2024. Discovery of the hidden world with large language models. *arXiv preprint arXiv:2402.03941*.

[31] 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*.

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

[33] 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*.

[34] 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.

[35] 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.

[36] Joakim Nivre. 2005. Dependency grammar and dependency parsing. *MSI report*, 5133(1959):1–32.

[37] 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.

[38] 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.

[39] Judea Pearl. 2009. *Causality*. Cambridge university press.

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

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

[42] 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*.

[43] 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).

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

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

[46] 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*.

[47] 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*.

[48] Masayuki Takayama, Tadahisa Okuda, Thong Pham, Tatsuyoshi Ikenoue, Shingo Fukuma, Shohei Shimizu, and Akiyoshi Sannai. 2024. Integrating large language models in causal discovery: A statistical causal approach. *arXiv preprint arXiv:2402.01454*.

[49] 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*.

[50] 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.

[51] 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.

[52] 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.

[53] 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*.

[54] 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.

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

[56] 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.

[57] 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.

[58] 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*.

[59] 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.

[60] Jiji Zhang. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896.

[61] 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.

[62] 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.