Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey

Xiaoyu Liu[1*], Paiheng Xu[1*], Junda Wu[2], Jiaxin Yuan[1], Yifan Yang[1], Yuhang Zhou[1], Fuxiao Liu[1], Tianrui Guan[1], Haoliang Wang[3], Tong Yu[3], Julian McAuley[2], Wei Ai[1], and Furong Huang[1]

[1] *University of Maryland, College Park*
[2] *University of California San Diego*
[3] *Adobe Research*
[*] *denotes equal contribution*

Abstract

Causal inference has shown potential in enhancing the predictive accuracy, fairness, robustness, and explainability of Natural Language Processing (NLP) models by capturing causal relationships among variables. The emergence of generative Large Language Models (LLMs) has significantly impacted various NLP domains, particularly through their advanced reasoning capabilities. This survey focuses on evaluating and improving LLMs from a causal view in the following areas: understanding and improving the LLMs' reasoning capacity, addressing fairness and safety issues in LLMs, complementing LLMs with explanations, and handling multimodality. Meanwhile, LLMs' strong reasoning capacities can in turn contribute to the field of causal inference by aiding causal relationship discovery and causal effect estimations. This review explores the interplay between causal inference frameworks and LLMs from both perspectives, emphasizing their collective potential to further the development of more advanced and equitable artificial intelligence systems.

1 Introduction

Recently Large Language Models (LLMs) have showcased remarkable versatility across a spectrum of critical tasks. An LLM is adept at tasks such as copywriting, enhancing original sentences with their distinct style and voice, responding to knowledge base queries, generating code, solving mathematical problems, and performing classification or generation tasks tailored to user requirements. Moreover, there has been a recent expansion into multi-modal variants, such as Large Visual Language Models (LVLMs) or Large Multi-modal Language Models, which broaden their input/output capabilities to encompass various modalities. This evolution has significantly enhanced both the potential and range of applications of these models.

In this survey, our primary focus is on Transformer-based Large Language Models (LLMs). The capability of LLMs is fundamentally rooted in their inference abilities, which dictates their proficiency in comprehending, processing, and providing solutions to various inquiries, as well as their adaptability to societally impactful domains. Consequently, extensive research efforts have been dedicated to measuring and enhancing these capabilities, ranging from assessing the reasoning abilities of LLMs to scrutinizing their decision-making processes and addressing challenges such as concept alignment across different modalities and mitigating hallucination. In addition, since LLMs are trained on extensive human knowledge with billions of parameters, they sometimes face challenges in appropriately prioritizing or downplaying what they have learned in different scenarios. This can lead to issues such as domain shift, where the model's performance degrades on data that differ from the training set, and long-tail bias, where infrequent examples are not handled as effectively.

In many instances, language tasks require not only predicting or generating text based on patterns in the data but also understanding the underlying causal mechanisms driving these patterns. Causal inference has shown great potential in improving predictive accuracy, fairness, robustness, and explainability of Natural Language Processing (NLP) models [23]. With the advent of generative LLMs, a significant transformation

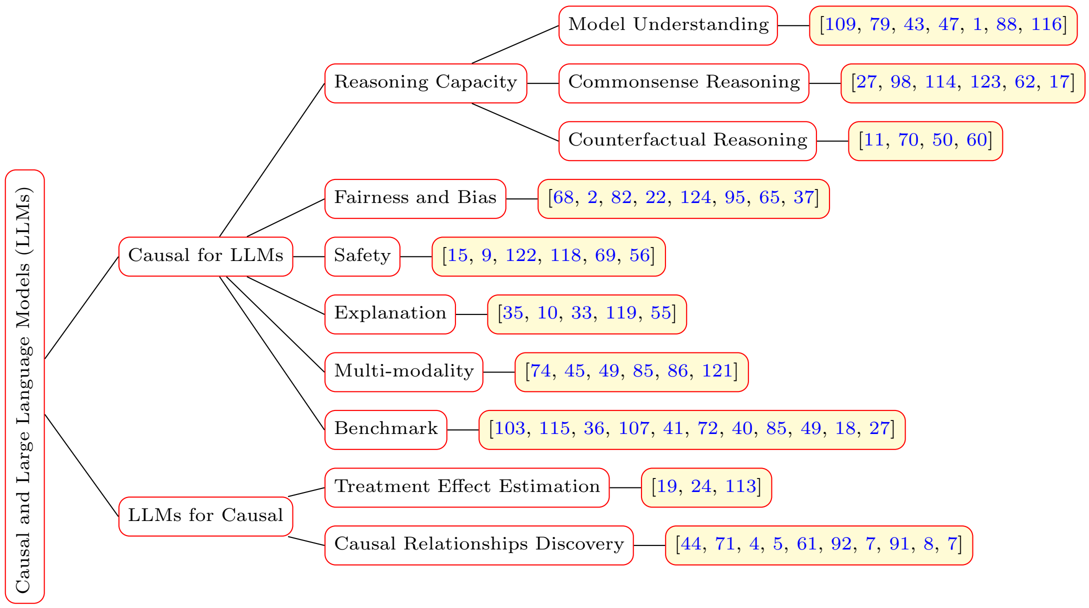

Figure 1: Structure Overview

has occurred across various NLP fields, attracting increased research interest in applying causal inference to address LLM-related challenges and augment their functionality. Such transformation also motivates this survey to outline causal methodologies and their implementation in LLMs, emphasizing their role in enriching our comprehension and application of language models.

Meanwhile, this survey also aims to explore how LLMs can help with the causal inference framework. Causal inference is formally defined as an intellectual discipline that considers the assumptions, study designs, and estimation strategies that allow researchers to draw causal conclusions based on data [75]. Causal inference has three main origins: potential outcomes, graphs, and structural equations, each serving unique purposes [110]. The potential outcomes framework [75] focuses on estimating causal effects through statistical inference and treatment comparisons. Graphical models, meanwhile, excel in mapping out causal pathways and visualizing relationships, with nodes representing variables and edges denoting directional influences. In this survey, we mainly discuss Pearl's formulation of causal graphs [76], which formalized causal graphical models for presenting conditional independence relations among random variables using directed acyclic graphs (DAGs).

We summarize how LLMs can help causal inference in its two important components, i.e., causal relationship discovery and treatment effect estimation. Determining causal relationships among variables is a fundamental step in a causal inference framework because the estimation of the causal effect of variable A on variable B requires causal assumptions on causal relationships of other variables with A and B. Traditionally, researchers rely on experts with subject matter knowledge to set the ground for such causal relationships. Causal discovery methods [108] provide an alternative by discovering causal graphs from observational data. LLMs have demonstrated abilities to determine such causal relationships based on pre-trained knowledge or given text. They can also be integrated with causal discovery methods to further increase the reliability of the outcome. Estimating treatment effects is central to causal inference but is hindered by the absence of counterfactual data in many cases. Leveraging the strong counterfactual reasoning abilities of LLMs, researchers have developed various ways to generate high-quality counterfactuals to enable treatment effect estimation.

The structure of the survey is given in Figure 1. We start with an introduction to recent advancements in Large Language Models in Section 2. Then we provide an overview of causal inference methods that are used and can be used to improve LLMs in Section 3. In the first half of the paper we go through how these methods are used in various problems in the LLM community: Section 4.1 overviews how causal methods are used to measure and improve the reasoning capacity of LLM, Section 4.2 and Section 4.3 focus on the fairness and safety issues while Section 4.4 introduces how the explainability of LLM are approached by causal inference methods. We also discuss the extension to construction and development of multi-modality large models in Section 4.5. Finally, we list existing work on evaluation and benchmarking LLMs from a causal perspective in Section 4.6. In the second half of the survey, we shift to how LLM extends the boundary of causal inference. Section 5.1 explains current assumptions, limitations and bottlenecks of causal inference. Section 5.3 and Section 5.2 states current works on improving treatment effect estimation and causal discoveries. We highlight several future directions in Section 6.

2 Background of Large Language Models

Large Language Models (LLMs) have transformed the way we interact with and process language, opening up new possibilities for natural language understanding, generation, and communication [34]. These models are in a constant state of evolution, continually expanding the limits of what is achievable in language processing and artificial intelligence [73]. In this paper, we mainly focused on Transformer-based LLMs, and we provide an overview of their recent progress in this session.

The major breakthrough of Large Language Models (LLM) came in 2017 when the Transformer [93] was introduced. In a groundbreaking shift, the Transformer technology emerged, mastering the art of grasping long-term linguistic connections. This innovation wasn't just theoretical—it allowed for simultaneous training on multiple GPUs, paving the way for the creation of significantly larger models [16]. Then came a pivotal moment in 2018: the unveiling of OpenAI's GPT-1 [78]. This was more than just another step in natural language processing (NLP); it was a leap, thanks to its transformer-based architecture. Boasting 117 million parameters, GPT-1 wasn't merely crunching data; it was weaving contextually coherent sentences, showcasing the transformative power of transformers in redefining NLP tasks [3]. Despite its initial limitations, GPT-1 didn't just make a mark; it laid the groundwork for a new wave of AI exploration and sparked intense competition in the realm of LLMs.

In 2020, OpenAI released GPT-3 [20], which was able to generate highly coherent and natural-sounding text [13]. It was a big deal because it showed just how awesome these huge language models could be at a bunch of different language tasks [26]. Riding on the high from GPT-3's success, OpenAI released the next iteration of their language model, GPT-4 [2, 67, 52] with the ability to generate even more coherent and natural-sounding text. Following GPT-4's success, Google wasn't far behind with their Bard [66]. Amazon jazzed up Alexa [42] with some cool AI features, Huawei jumped in with their Pangu models [111] and Alibaba proposed their QWEN models [6]. Then Meta brought out this thing called LLaMA [90], which was all about the first open-source foundation models. Compared to LLaMA, LLaMA2 [90] has made more explorations in reinforcement learning from human feedback (RLHF) and developed a chat-oriented version called LLaMA-chat, which generally outperforms existing open-source models across a range of helpfulness and safety benchmarks. A large number of researchers have extended LLaMA models by either instruction tuning [57] or continual pretraining. Alpaca [89] is the first open instruct-following model fine-tuned based on LLaMA. In addition, Vicuna [77] is another popular LLaMA variant, trained upon user-shared conversations collected from ShareGPT [31].

As diverse versions of LLMs emerge, they encounter common challenges. In this survey, we demonstrate that many of these challenges can be effectively addressed through causal methods. These methods encompass enhancing the reasoning capacity of LLMs, tackling fairness concerns and mitigating potential biases, ensuring safety, and enhancing the explainability of model outputs, and their extension to multi-modality versions.

Building upon this progress, there is now a burgeoning interest in expanding the scope of these models to encompass visual data, giving rise to the emergence of Large Visual Language Models (LVLMs). These models aim to integrate the understanding of both textual and visual information, opening up new avenues for more comprehensive AI systems capable of interpreting and generating content in multimodal formats.

One of the most prevalent approaches is inserting visual features as supplementary inputs to LLMs and aligning them with textual features. This method has been adapted in several large vision-language models (LVLMs) such as MiniGPT-4 [127], LLaVA [59], Mplug-Owl [106] and so on [55, 57, 117, 96, 106, 58, 51]. In this survey, we also demonstrate that the causal methods could help concur challenges encountered in existing LVLMs as above.

3 Brief Introduction of Causal Inference

In this section, we present the background knowledge of causal inference, including task descriptions, basic concepts and notations, and general solutions.

Generally speaking, the task of causal inference is to estimate the causal relationship among variables. The variables of interest are referred to as *treatment*, naturally, the effects of treatments are referred to as *treatment effects*. For example, suppose two treatments can be applied to patients: Treatment Plan A and B. When A is applied to a certain patient cohort, the recovery rate is 70% while when B is applied to *the same cohort*, the recovery rate is 80%. The change of recovery rate is the effect of that treatment assets on the recovery rate.

Ideally, the treatment effect can be measured as follows: applying different treatments to the same cohort, and then the difference in the effect is the treatment effect. However, in real-world scenarios, this ideal situation because it is impracticable for perfectly controlled experiments in most cases. For example, in the above case, you can only apply one treatment to the same cohort at the same time. In reality, an alternative is to conduct random controlled trials, in which the treatment assignment is controlled, such as a completely random assignment. In this way, the groups receiving different treatments can be used to measure the difference in effect. Unfortunately, even performing randomized experiments is expensive, time-consuming, and may cause ethical concerns in some cases. Therefore, estimating the treatment effect from observational data has attracted growing attention due to the wide availability of observational data, and methods are developed for the investigation of the causal effect of a certain treatment without performing randomized experiments.

3.1 Potential Outcome Framework

One of the most influential frameworks in identifying and quantifying causal effects in observational data is the potential outcomes framework [80]. The potential outcomes approach associates causality with manipulation applied to *units*, and compares causal effects of different treatments via their corresponding potential outcomes. Following [80], we state basic concepts in the potential outcome framework.

Unit. A unit is the atomic research object in the treatment effect study. A unit can be a physical object, a firm, a patient, a person, or a collection of objects or persons, such as a classroom or a market, at a particular time point [80]. Under the potential outcome framework, the atomic research objects at different time points are different units.

Treatment. Treatment refers to the action that applies (exposes, or subjects) to a unit. For each unit-treatment pair, the outcome of that treatment when applied to that unit is the **potential outcome**.With N treatments $T = \{1, 2, 3, ..., N\}$, the potential outcome of applying treatment T_i is denoted as $Y(T = T_i)$. The **observed outcome** is the outcome of the treatment that is actually applied. And the **counterfactual outcome** is the outcome if the unit had taken another treatment.

Treatment Effect The treatment effect can be quantitatively defined using the above definitions. The treatment effect can be measured at the population, treated group, subgroup, and individual levels. At the population level, the treatment effect is estimated as the Average Treatment Effect (ATE). At the subgroup level, the treatment effect is called the Conditional Average Treatment Effect (CATE).

Definition 3.1 (Binary Average Treatment Effect(ATE))**.** Suppose we want to measure the treatment effect of a treatment $T = 1$. Then the average treatment effect is defined as:

$$\mathbb{E}[Y(T = 1) - Y(T = 0)] \tag{1}$$

where $Y(T = 1)$ and $Y(T = 0)$ denote the potential treated and control outcome of the whole population respectively.

Definition 3.2 (Conditional Average Treatment Effect (CATE))**.**

$$\mathbb{E}[Y(T = 1)|X = x] - \mathbb{E}[Y(T = 0)|X = x] \tag{2}$$

where $\mathbb{E}[Y(T = 1)|X = x]$, $\mathbb{E}[Y(T = 0)|X = x]$ are the potential treated and control outcome of the subgroup with $X = x$.

At the individual level, the treatment effect is defined as Individual Treatment Effect (ITE). In some literature, ITE is treated as the same as CATE [75].

3.2 Causal Graphical Models

The potential outcome framework is powerful in recovering the effect of causes. In a potential outcome framework, causal effects are answered by specific manipulation of treatments. However, when it comes to identifying the causal pathway or visualizing causal networks, the potential outcome model has its limitations. In the front of the challenge, causal graphical models utilize directed edges to represent causalities and encode conditional independence among variables in the graphs.

3.2.1 Structural Equation Models (SEMs)

One of the most widely-spread formulations is the Structural Equation Model [99, 76], where linear structural equation models are used to present causal relationships by directed edges, which differentiate correlation from causation when the graph structure is given. The linearity assumption was later been relaxed by [76] and it formalized causal graphical models for presenting causal relations using Directed Acyclic Graphs (DAGs).

Specifically, consider the random variable $\mathbf{X} \in \mathcal{R}^{D\times N} = [X_1, X_2, ..., X_N]$, the linear SEM consists of a set of equations of the form:

$$X_i = \beta_{0i} + \sum_{j\in pa(X_i)} \beta_{ji} X_j + \epsilon_i, \quad i = 1, 2, 3, ..., N \tag{3}$$

where $pa(X_i)$ denotes the set of variables that are direct parents of X_i. $\epsilon_1, \epsilon_2, ..., \epsilon_N$ are mutually independent noise terms with zero mean, β_{ji} are coefficients that quantify the causal effect of X_j on X_i.

While the non-parametric SEM takes the form:

$$X_i = f_i(\mathbf{X}_{pa(i)}, \epsilon_i), \quad i = 1, 2, 3, ..., N \tag{4}$$

The random variables $\mathbf{X}$ that satisfies the model structure of the form in Equation (3) or Equation (4) can be represented by a directed acyclic graph (DAG) $G = (V, E)$, where V is the set of associated vertices, each corresponding to one of a variable of interest X_i, and E is the corresponding edge set.

With pre-specified DAG and assumptions on the latent variables, the coefficients between the latent variables are identifiable[46].

3.2.2 Bayesian Network

Causal inference can be naturally embedded in graphical model frameworks since the dependencies and interactions between variables can be presented by graphs with probabilistic distributions, in which nodes correspond to variables of interest and edges represent associations. One general solution except for SEMs is to use a Bayesian Network to represent the causal relationship.

In Bayesian networks, causalities among variables are represented in the form of DAGs with directed edges carrying causal information.

A joint probability distribution $\mathbb{P}$ factorizes with respect to a DAG $\mathcal{G}$ if it satisfies:

$$f(X_1, X_2, ..., X_N) = \prod_i f(X_i|\mathbf{X}_{pa(i)}) \tag{5}$$

In the next section, we show a comprehensive survey of how existing works help with the tasks and challenges in LLMs in detail.

4 Causal Inference for Large Language Models

LLMs can significantly benefit from causal inference as it enhances their ability to understand and reason about cause-and-effect relationships within data. In this section, we review how LLMs can benefit from a causal view in the following perspectives, understanding and improving the LLMs' reasoning capacity (Section 4.1), addressing fairness (Section 4.2) and safety (Section 4.3) issues in LLMs, complementing LLMs with explanations (Section 4.4), and handling multimodality (Section 4.5). We then organize benchmark datasets from these perspectives in Section 4.6.

4.1 Reasoning Capacity

4.1.1 Model Understanding

LLMs have demonstrated many emerging abilities in language generation and certain reasoning tasks [14, 44]. As the reasoning process is often associated with causal factors, it is logical to first understand and evaluate LLMs' reasoning ability from a causal lens. Zečević et al. [109] argued LLMs are not causal and hypothesized that LLMs are simply trained on the data, in which causal knowledge is embedded. Thus, in the inference stage, the LLMs can directly recite the causal knowledge without understanding the true causality in the context. Similar behaviors are exhibited in a Causal Reasoning Assessment Benchmark, CRAB, that consists of 1.8K causal frames and 352 causal chains in real-world narratives [79], where LLMs are required to output the causality class (high, medium, low, or no causality) between variables. They show that LLMs can capture explicit causal statements in pre-training data, but face performance drop when applying causal reasoning to new distributions, i.e., events that happened after the pre-training phase. Kim et al. [43] examined LLMs' abilities in understanding the causalities of both scientific papers and newspapers. Their evaluation protocol is designed for the LLM to tell whether a statement is *causal*, *conditional causal*, *correlational*, or *no relationship*. The results show that ChatGPT performs worse than a fine-tuned BERT model in understanding causality. Abdali et al. [1] show the effectiveness of applying LLMs to diagnose the cause of issues from Microsoft Windows Feedback Hub. Li et al. [47] showed that LLMs can identify dynamical (spatio-temporal) effects. However, how to infer the relationship and interactions of them is still challenging for LLMs, which are more emphasized as causal structures in causal inference.

Another important line of work is to understand LLMs' hallucination and faithfulness in knowledge reasoning by considering causal effects. Tang et al. [88] proposed a multi-agent system, CaCo-CoT, where some LLMs are *reasoners* and others are *evaluators*. *Reasoners* try to provide causal solutions, while *evaluators* try to challenge the *reasoners* with counterfactual candidates. With the cooperative reasoning framework, CaCo-CoT helps to improve causal-consistency. Zhang et al. [116] identified the potential knowledge bias pretrained in the LLMs as the confounder which causes incorrect answers and hallucinations. Zhang et al. [116] proposed a chain-of-question framework to generate sub-questions necessary to answer a question, and involve humans in the loop to provide the correct answers. With human annotation in the loop, the confounding causal effect can be reduced, thus mitigating the spurious correlation.

4.1.2 Commonsense Reasoning

Commonsense reasoning involves the ability to apply everyday knowledge and intuitive understandings of the world to make decisions or draw conclusions, which is vital for LLMs' contextual understanding and human-like interactions [21, 84]. This section briefly summarizes the commonsense reasoning ability of LLMs under various settings [27, 98] and the employment of causally motivated methods in improving commonsense causality reasoning [114, 123, 62, 17, 120].

Through the evaluation of ChatGPT's performance on event causality identification, causal discovery, and causal explanation generation, Gao et al. [27] have shown that ChatGPT is not a good reasoner but a good causal explainer. ChatGPT, and even gpt-4 is outperformed by baseline methods based on fine-tuned small pre-trained language models on event causality identification and does poorly on causal discovery. They also observe serious hallucinations on causal reasoning under In-Context learning and Chain-of-Thought settings A similar conclusion has also been obtained in [98], which analyzed the causal question answering capabilities of LLMs. It concluded that LLMs did not arrive at their answers through reasoning but through memorization of the corresponding question and answer pair. Even if ChatGPT may not have causal reasoning ability, it generates accurate and detailed causal explanations in some cases [27].

To improve the commonsense causality reasoning of LLMs that identify causes from effects in natural language, ROCK (Reason O(A)bout Commonsense K(C)causality) [114] balances confounding effects using temporal propensities. The central question in this framework is the estimation of Average treatment effect (ATE). Given two events such that E_1 precedes E_2, the strength of causation from E_1 to E_2 can be estimated by the change in the likelihood of E_2's occurrence due to intervening on E_1, denoted as ATE. While ROCK adopts a potential-outcome framework, Chen et al. [17] uses a conversation cognitive model based on intuition theories, and transforms intuitive reasoning into a structural causal model. Aiming to improve conversation reasoning, the authors incorporate mental states such as desires, memory, experience and emotion, as an unobservable exogenous variable that implicitly affects the corresponding observable utterance:

$$H = AH + E, \tag{6}$$

where H is the embedding of the utterance, E is the exogenous variable, and A is the adjacency matrix.

Other than facilitating reasoning ability of LLMs directly, Zheng et al. [123] use causal inference to preserve commonsense knowledge from Pre-trained language models for fine-tuning to prevent catastrophic forgetting; Lu et al. [62] focus on improving LLM's ability in generalized procedural planning with commonsense-infused prompts. In particular, Zheng et al. [123] abstract the fine-tuning process as a causal graph and discovered that catastrophic forgetting is due to missing causal effects from the pre-trained data. To preserve old knowledge from a pre-trained language model, an objective for fine-tuning (causal effect tuning) is introduced. For procedural planning tasks, Lu et al. [62] proposed to learn causeeffect relations among complex goals and stepwise tasks, and reduced spurious correlation among them via front door adjustment.

4.1.3 Counterfactual Reasoning

Another potential use of LLMs' generative capacity is to generate counterfactuals for data-augmentations for small language models. Given a text x and a black-box classifier B, the counterfactual text $\tilde{x}$ of text x should satisfy the following requirement [11, 70]:

1. $\tilde{x}$ has a different class than x, $B(x) \neq B(\tilde{x})$.
2. x and $\tilde{x}$ differ only by minimal lexical changes.
3. $\tilde{x}$ is a feasible text and the commonsense constraint is satisfied.

This section discusses briefly the performance of LLMs in generating counterfactuals [50, 60] and the effort of improving their qualities [70].

Li et al. in [50] examined the effectiveness of LLMs' performance in generating counterfactuals under four tasks: (1) sentiment analysis (SA) that alters sentiment polarity; (2) natural language inference (NLI) that given a premise sentence and a hypothesis sentence, make a change to the hypothesis sentence to alter the relationship between it and the premise sentence; (3) named entity recognition (NER) that changes the entities in a sentence whose type is the same as the original entity type; (4) relation extraction (RE) that changes the relation between the head and tail entity. It has been demonstrated that for simple tasks like SA and NLI, data augmented via LLMs can mitigate potential spurious associations. For more complicated tasks like RE, LLMs may generate low-quality counterfactuals. In order to generate high-quality counterfactual, detailed instructions are crucial. However, counter-intuitively, chain-of-thought doesn't always help as it may even lead to significant performance decreases under some settings. On the other hand, Liu et al. in [60] evaluated abductive reasoning and counterfactual reasoning abilities for large language models of code (Code-LLMs) and compared them with text models. With code-prompts designed to tackle causal reasoning tasks, it has been shown in [60] that Code-LLMs achieve better results than text models.

As shown in the evaluation [50], though LLMs as counterfactual generators can enhance the performance of small language models on simple tasks such as sentiment analysis and natural language inference, the generated counterfactuals fail to have any significant effect on complex tasks like relation extraction. Miao et al. in [70] claim that this is due to the failure of identifying causal terms correctly and ignoring the commonsense constraint. To amend this, they proposed a framework to generate commonsense counterfactuals for stable relation extraction via an intervention-based strategy. This framework is demonstrated to have enhanced the stability on relation extraction tasks under various settings including the low-resource, out-of-domain and adversarial-attack scenarios.

4.2 Fairness and Bias

Fairness and bias are pivotal factors in deploying language models effectively and ethically. Bias is common in pretrained language models as they capture and potentially amplify undesired social stereotypes and biases [68, 2, 126, 100, 96]. An example of bias in language models includes gender associations with specific professions, such as male firefighters and female nurses [82]. Causality-based methodologies offer a promising approach for mitigating biases in language models by discerning the origins of bias through a causal perspective. Bias mitigation is then followed by eliminating the unwanted spurious correlation between generative factors though different types of causal intervention or causal invariant learning.

Ding et al. [22] introduce a proxy variable related to gender bias in the causal graph, and use two different ways to eliminate the potential proxy bias and unresolved bias under the linear structural equation model. Zhou et al. [124] believe that the backdoor path between the ground truth label and the non-causal factors is the source of bias, and uses the Independent Causal Mechanism (ICM) principle to mitigate bias. Their proposed method, Causal-Debias, achieves causal intervention and creates interventional distributions with respect to different demographic groups by augmenting and expanding the original data distribution. Wang et al. [95] raise concerns over the precision of parameter estimation in existing causal models and introduce several intermediate variables that are the causal children of raw text and the causal parent of the input to language models. Under this assumption, they propose to eliminate bias by performing a 'do' operation on the intermediate variables for both white-box LLMs and black-box LLMs. Madhavan et al. [65] consider the tokens generated by generative language models trained with causal language modeling (CLM) objectives as a causal graph, and analyze the bias under this model. Jenny et al. [37] use Activity Dependency Networks (ADNs) to describe the causality effect between normative variables, such as clarity and authenticity, to structure the cause of bias. The authors argue that using ADNs can better explain previously simplified views of bias using just correlation, and display the complexity nature in identifying and mitigating biases in large language models.

4.3 Safety

With the application of language models in multiple downstream tasks, researchers have observed the unreliability phenomenon of LLMs in knowledge probing [15] or downstream inference tasks [9, 122, 118, 69]. There is increasing interest in applying the causal inference technique to analyze the causality of the non-robustness of the model and adjust the treatment to resolve the challenges [15, 118, 122, 56].

Knowledge probing To elicit the knowledge encoded in LLMs, previous work uses prompt-based prompting, that is, querying LMs with task-specific prompts. However, in this process, LLMs face challenges of unreliability [15, 102], such as using shortcuts to complete the probing and generating different predictions for the semantically equivalent prompt. To explore the reason behind the non-robustness, an empirical study from [15] constructs a structural causal model (SCM) containing 11 variables and identifies three types of bias in the knowledge probing procedure: prompt preference bias, instance verbalization bias, and sample disparity bias. Blocking the corresponding backdoor path for each type of bias effectively eliminates the bias. This finding reveals the potential of constructing an SCM with LLM-related variables to improve LLM robustness.

Downstream Inference Tasks In addition to knowledge probing, LLMs show vulnerability when encountering attacks on the prompts in downstream inference tasks [118, 122]. By simply translating text tokens in input prompts to emoji sequences, LLMs generate more severe hallucinations [122]. Some neurons within LLMs also have an unreasonably high causal effect on the model response and, by changing the value of that neuron, LLMs will produce meaningless responses [122].

The main reason why LLMs fail in prompt attacks is that it uses the spurious correlation to make an inference [118]. Training LLMs to learn the causal relationship between input x and output y is an intuitive method of better resisting prompt attacks. The randomized smoothing technique [112, 39] can model the interventional distribution $p(y|do(x))$ by assuming discrete adversarial perturbations as the Gaussian distribution [118]. The method of smoothing in the latent semantic space is more robust against known attacks such as word substitutions, paraphrasing, and token position change [118].

Causal inference techniques are useful to induce robustness for LLM applications. For the setting without prompt attacks, mitigating LLM unreliability can be achieved by identifying the causality and blocking the backdoor path. Moreover, smoothing the latent semantic space is effective in resisting prompt attacks. Despite the progress of causality analysis, there are some directions to explore, such as the causal relationship between the pretraining corpus and model robustness and the evaluation method of robustness in the generation task.

4.4 Explainability

Explainability in LLMs refers to the capacity to elucidate how these models arrive at their conclusions, enhancing transparency and trustworthiness in AI decision-making processes [32, 54]. Many work have tried to explain and understand the inner workings of LLMs [35, 10, 33, 119, 55]. We summarize research efforts that probe the causal mechanism in LLMs from the following three directions: intervening the inputs or prompts, intervening inner components of LLMs, and abstracting the working mechanism into a causal graph.

Inputs or Prompt Intervention Input intervention, as a data-centric method, is to create counterfactual input text by changing the treated feature in the text and to observe the model behavior on the original and counterfactual texts. Recruiting individuals to produce counterfactual texts typically incurs significant expenses. However, the advent of LLMs has shown that generating counterfactual inputs can be achieved with less cost. LLMs can first identify the features in input texts causally associated with the predictions and are capable of changing the identified features to create the counterfactual texts [12, 28]. These counterfactual texts can be utilized to investigate the causality of the LLM and can serve as a training dataset to learn a matching model, where the matched counterfactual pairs have similar embeddings [28].

Various works have developed different prompting methods and found whether the prompting methods are causally associated with the final output of LLMs [81, 125, 97]. However, the causal effect of prompting methods, such as chain-of-thought (CoT), and the final output is ambiguous. Prompt intervention, which alters only one particular aspect of prompts, is proposed to understand the contributions of each component of prompts on model behavior [64]. The experiments from [64, 38] first find that the linguistic features and the grammar have a large effect on the LLM outputs. Then, intervention in intermediate variables in prompts leads to consistent final answers with the expected output of the hypothesized causal model [87]. These findings suggest that LLMs realize the causal model suggested by their CoTs to a high extent, but LLMs also utilize spurious correlations such as sentence length to make responses.

Input or prompt interventions are data-centric methods to probe the LLM mechanism, applied in both open-source or black-box LLMs. However, the detailed information cascade inside the LLMs cannot be discovered by such methods, so intervention on the inner model components is proposed.

Inner Component Intervention The attention mechanism and multilayer perceptron (MLP) layers are the essential components in the structure of state-of-the-art (SoTA) LLMs. Existing work exchanges the activation values in MLP and attention layers of different inputs to probe the function of MLPs and attention mechanism in generating the answers [83]. The experimental results indicate that LLMs transfer the information of inputs from midsequence early layers to the final token using the attention mechanism. Due to the complexity of LLMs, current work focuses only on math word problems with four fundamental arithmetic operators [83]. It is an interesting direction to generalize the component intervention to other downstream tasks.

Causal Graph Abstraction An intuitive way to characterize causality within LLMs is to abstract the working mechanism of LLMs into a causal graph. Boundless Distributed Alignment Search (DAS) [101] by replacing brute-force searching the original DAS [29] with learnable parameters, has been effective on the Alpaca model [89]. Given four pre-defined causal models, Boundless DAS extracts two of them as the accurate hypotheses as the abstracted causal graph of the Alpaca model. However, the Boundless DAS method is restricted by the given causal hypothesis, and the future direction can explore how to abstract the causal graph in LLMs without prior causal graphs.

Current explainability work from a causal view use LLMs to interpret the causality in real-world events. LLMs can generate high quality counterfactuals. By altering the inner values of LLMs and abstracting the causality in LLMs, the current work has pointed out a direction to characterize the inner causality of LLMs on various tasks.

However, with these mentioned works in LLM interpretation, the scope of interpretation focuses on tasks that have a clear causal graph between task inputs and outputs such as simple math word problems. Probing the causality in LLMs for more complex generation tasks, such as question answering or summarization tasks, could provide more insight on the inner mechanism of LLMs.

4.5 Multi-modality

Large vision-language models have been popular in many applications [63, 48, 59, 30]. How to conduct causal reasoning on both images and texts can be crucial in correctly answer multimodal questions. Pawlowski et al. [74] examined LLMs' causal reasoning abilities and showed that the causal knowledge in the language models can be too strong a prior which often causes the model to neglect visual evidence. Ko et al. [45] proposed to alleviate the problem by adding self-consistent generation prediction, in which the three input V, Q and A are individually predicted based on the other two inputs. Specifically, Li et al. [49] proposed a image generation framework with causal reasoning and created a novel VQA datasets whose questions requires causal explanations.

Another important question is to understand the spatial-temporal causal relationship of the visual elements within the images and videos. Su et al. [85] proposed a framework, CaKE-LM, to leverage the pretrained causal knowledge in the language model to understand the causal relations of the events in a video. Based on the generated causal reasoning results, CaKE-LM can further generate the question-answer pairs and construct a new benchmark for causal video question-answering. Tai et al. [86] proposed the link-context learning method to strengthen the LLM's in-context learning abilities by instructing the model to understand the underlying causal relationship between the demonstration data points. Zhao et al. [121] suggested that there are two types of causal relationships in VQA. They proposed a prompting method, causal context generation, to engage contextual information for better VQA precision.

4.6 Evaluation and Benchmark

In this section, we list existing evaluation metrics and benchmarks from a causal perspective for LLMs as listed in Section 4.6. The causal evaluation mainly focuses on three aspects: the Model Understanding (MU) ability, the Commonsense Reasoning (CR) ability, the Counterfactual Reasoning (CF) ability, and the Fairness/Debias (FD) ability.

Reference	**MU**	**CR**	**CF**	**FD**	**Language**	**Multimodal**
ECHo [103]	✓	✓				✓
CREPE [115]		✓			✓	
CLOMO [36]			✓		✓	
IfQA [107]			✓		✓	
Cladder [41]	✓	✓			✓	
MoCa [72]	✓				✓	
CORR2CAUSE [40]	✓				✓	
CVidQA [85]	✓					✓
VQAI [49]	✓					✓
Chen et al. [18]			✓		✓	
Gao et al. [27]	✓				✓	
CRAB [79]	✓				✓	
HELM [53]				✓	✓	
Fair-Prism [25]				✓	✓	
Biasasker [94]				✓	✓	

Table 1: We summarize the existing evaluation benchmarks. Based on the evaluation tasks, we categorize the benchmarks into three categories: Model Understanding (**MU**), Commonsense Reasoning (**CR**), Counterfactual Reasoning (**CF**) and Fairness/Debias (**FD**). Based on the modalities of the data samples, we identify the benchmarks with only textual inputs (**Language**) and those with multimodal inputs (**Multimodal**).

The benchmarks in model understanding (MU) focus on evaluating and understanding existing LLMs' causal reasoning abilities in both natural language [41, 72, 40, 27, 79] and vison-language [85, 49, 103]. In addition, some benchmarks [72, 27] also provide model understanding in comparison with human causal reasoning and moral judgments. Commonsense reasoning benchmarks (CR) evaluate LLMs on tasks that require extensive commonsense knowledge for both textual-only context [115, 41] and multimodal context [103]. Contexts with commonsensical and anti-commonsensical are constructed in [41], to further investigate whether LLMs use averaged-out causal reasoning. Evaluating LLMs' counterfactual reasoning (CF) abilities is essential in enabling explainable model reasoning and calibration of the generated rationales. Huang et al. [36] introduce a specific task and benchmark for assessing LLMs' logical counterfactual thinking. Yu et al. [107] contribute a novel dataset to challenge LLMs in counterfactual reasoning in an open-domain QA context. Chen et al. [18] investigate the ability of LLMs to provide explanations that aid in understanding their reasoning process, particularly in the context of counterfactual scenarios. The fairness and bias (FD) evaluations are particularly in addressing biases, fairness, and the overall transparency of language models. HELM [53] is a comprehensive evaluation benchmark including previously neglected areas for fairness. Fair-Prism [25] focuses specifically on fairness-related harms in models, which are identified and measured by detailed human annotations. Biasasker [94] presents an automated framework to identify and measure social biases by probing the models with specially designed questions.

5 Large Language Model for Causal Inference

Causal inference, serving as a potent tool for addressing challenges in LLMs, heavily relies on world knowledge. As previously mentioned, there are three primary origins of causal inference: the potential outcome framework, graph-based causal methods, and the structural equations community. The potential outcome framework relies significantly on several assumptions to facilitate the comparison of treatment effects among groups/individuals. One of the most challenging aspects of applying the potential outcome framework lies in ensuring that these assumptions hold true in reality. In this section, we first examine these assumptions and subsequently illustrate how they are relaxed in existing literature. The graph-based causal methods and structural equation models also necessitate a certain level of understanding of the underlying causal graph. For instance, Directed Acyclic Graphs (DAGs) serve as a fundamental assumption, and many structural equation models assume a degree of linearity or that the input distribution adheres to specific probability distributions. In our review, we also explore how existing methods verify the distribution in the input data and extend current methodologies to accommodate more complex distributions with the assistance of LLMs.

5.1 Fundamental Assumptions in Estimating Treatment Effect

In existing causal inference literature, several assumptions are adopted to estimate the treatment effect. Here we discuss the three most commonly used assumptions and then show how the development of large language models could help relax or challenge these fundamental assumptions.

Assumption 5.1 (**Stable Unit Treatment Value Assumption**)**.** The potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which leads to different potential outcomes.

This assumption emphasizes the independence of each unit when estimating the treatment effect that the units do not interact with each other. From a statistical perspective, it is equivalent to assuming each treatment assignment subject is iid.

Assumption 5.2 (**Ignorability/Unconfoundedness**)**.** Given the observable background variable, X, treatment assignment T is independent of the potential outcomes.

The assumption 5.2 states that if the background variable X is the same for two patients, then (1) the treatment assignment should be the same. (2) the potential outcome should also be the same. In other words, these two patients are treated as identical units, thus can be used to estimate the treatment effect if they are assigned different treatments in the static dataset, as the treatment assignment is treated as random.

Assumption 5.3 (**Positivity**)**.** For any value of X, treatment assignment is not deterministic. i.e.,

$$P(T = t|X = x) > 0, \forall t, x \tag{7}$$

This assumption tries to guarantee that the treatment effect can be estimated and that we can always find a comparable sample. In the binary case, if the task is to estimate the performance of a specific treatment $T = 1$, we would need to compare the potential outcome of patients receiving the treatment against those who are not treated, requiring dataset points in both cases.

5.2 Treatment Effect Estimation

One main obstacle of traditional causal methods is the lack of counterfactual data, making the estimation of causal effects a difficult problem in practice. Chen et al. [19] proposed a new method for automatically generating high-quality counterfactual data at scale called DISCO (DIStilled COunterfactual Data). Specifically, it prompts to generate of phrasal perturbations with a large general language model. Then, a task-specific teacher model filters these generations to distill high-quality counterfactual data. In addition, Feder et al. [24] apply treatment effect estimation to align knowledge for generalization towards different domains. Zhang et al. [113] try to optimize the treatment effect estimation on unlabeled datasets by performing self-supervised causal learning through LLMs. Through exploring the primal-dual connection between optimal covariate balancing and self-attention, their method facilitates zero-shot causal inference through the final layer of a trained transformer-type architecture, contributing to a foundation model for treatment effect estimation.

5.3 Causal Relationships Discovery

Discovering causal relationships between variables is a fundamental step in causal inference as it enables the identification and estimation of causal effects. In this section, we introduce papers discussing how LLMs can help discover causal relationships.

One line of this work focuses on casual relationship extraction or causality extraction which extracts causal relationships between two variables from text directly. Traditional methods rely on linguistic cues such as causal connectives (e.g., "cause", "because", and "lead to") and grammatical patterns to identify causal pairs [104]. A later work utilizes the power of statistical machine learning and deep learning to tackle this task in a supervised learning setting [105]. As LLMs show promising potential with reasoning capacities as introduced in Section 4.1, many works use LLMs as a query tool to determine the edge direction between two given variables. For example, Kıcıman et al. [44] show that LLMs can achieve competitive performance in determining such pairwise causal relationships with accuracies up to 97%. Analyses in the medical domain [71, 4, 5] exhibit similar observations. However, other studies highlight LLMs' limitations of such pairwise causal relationships. For example, sensitivity to prompt design leads to inconsistent results [61]; pairwise judgments can lead to cycles in the full causal graph [92]; pairwise judgments require large computational cost when applying to a large-scale dataset, N variables would require $\binom{N}{2}$ prompts [7]; LLMs still provide false information despite achieving strong results in most cases [61, 91]. Long et al. [61] propose strategies for amending LLMs' output based on consistency properties in causal inference [61].

To alleviate the impact of erroneous causal information from LLMs, previous works have integrated LLMs with traditional causal discovery methods. Causal discovery or causal structure learning is the task of recovering causal graphs from observational data whenever possible [108]. Traditional causal discovery methods mainly include constrained-based methods that exploit a sequence of statistical tests and score-based methods that structure around the maximization of the fitness of a graph through a space of possible graphs. Vashishtha et al. propose two algorithms that combine LLMs with causal discovery methods: the first uses causal order from an LLM to orient the undirected edges outputted by a constraint-based algorithm and the second utilizes the LLM causal order as a prior for a score-based algorithm [92]. Ban et al. incorporate LLM-driven causal statements as qualitative ancestral constraints in the Bayesian network structure to guide data-driven algorithms [8], which benefits smaller-scale problems, but encounters difficulties with larger datasets due to inaccuracies in the LLM-derived constraints. They then propose an iterative framework that employs LLMs to validate the accuracy of edges in the learned causal graph and fine-tune the causal discovery process based on LLM feedback [7].

6 Future Directions

Theoretical understanding of LLM's reasoning capacity. Causal inference methods offer a promising avenue for gaining deeper insights into the reasoning capacity of Large Language Models (LLMs). A potential approach involves employing treatment effect estimation techniques to assess their performance on specific tasks. The gold standard for treatment effect estimation is comparing potential outcomes across various treatments in a controlled experiment. However, the practical challenges of conducting such experiments in real-world scenarios make the interactive nature of LLMs an ideal candidate for investigation.

Leveraging their inherent interactivity, researchers could explore quasi-experimental settings, utilizing the natural variations in responses to infer causal relationships. It is crucial to acknowledge the complexities associated with interpreting LLMs, potential biases in training data, and the intricate nature of language understanding tasks. Additionally, considering ethical implications and addressing biases in both training and evaluation data is essential when delving into the reasoning capacity of LLMs using causal inference methods.

Efficient training and inference in LLMs. As model scales and training data continue to expand, the process of knowledge updating and inference in Large Language Models (LLMs) becomes increasingly resource-intensive. Consequently, it becomes imperative for us to devise methods that can efficiently and judiciously update the knowledge base of pre-trained models. In this context, causal inference methods can play a crucial role by offering guidance in quantifying efficiency. By establishing causal relationships between methods of interest and the existing knowledge collection, these methods can help assess the impact of different updating strategies. This approach not only addresses the resource challenges associated with knowledge updating but also contributes to a more nuanced understanding of the evolving knowledge landscape in LLMs.

LLM-based counterfactual estimation and augmentation. As a general expert, Large Language Models (LLMs) can significantly contribute to overcoming the current limitations of causal inference methods. A common assumption in many causal methods, as indicated in Assumption 5.3, is the existence of corresponding data points for every treatment. However, this assumption often proves untrue, particularly when dealing with imbalanced minority data that may not support meaningful learning. LLMs, functioning as versatile experts, have the potential to address this challenge by aiding in data augmentation for minority data. Through their comprehensive understanding of language and context, LLMs can enhance the availability of diverse data points, facilitating more robust and effective causal inference in situations where traditional methods may struggle due to data imbalances. Similarly, many methods operate under the assumption of unconfoundedness (Assumption 5.2) within the potential outcome framework, a condition considered quite strong. Historically, this assumption has been accepted due to a lack of domain knowledge regarding the underlying causal graph or identification of potential confounders. However, with the advent of Large Language Models (LLMs), there is an opportunity to alleviate this stringent limitation. LLMs can act as general experts, offering valuable information about potential causal graphs and knowledge. This transformative capability of LLMs opens avenues to enhance our understanding of causal relationships, addressing a historical challenge where unconfoundedness assumptions were made due to limited domain knowledge.

7 Conclusion

At its core, a large language model (LLM) is like a vast library of knowledge. One of the ongoing challenges is figuring out how to extract and use this knowledge effectively. The key to improving LLMs lies in enhancing their ability to understand cause and effect – essentially, how things are connected. Causal reasoning is crucial for making LLMs smarter. Looking at it from a causal inference perspective, we find a valuable framework that helps boost the effectiveness of LLMs. Meanwhile, as keepers of human knowledge, LLMs can even help overcome limitations in causal inference by providing broad expertise that goes beyond existing constraints, reshaping our understanding in this important area and bringing new vitality to this area.

In this survey, we offer a thorough examination of the current landscape where large language models (LLM) intersect with causal inference. We delve into how causal inference contributes to LLM, enhancing aspects such as reasoning, fairness, and safety, along with the explainability of LLM. Additionally, we explore how LLM, in turn, broadens the horizons of causal inference. Across these categories, our survey provides in-depth discussions, comparisons, and concise summaries of the methods scrutinized, offering a comprehensive overview of the current state of research at this intersection. The available benchmark datasets and open-source codes of those methods are also listed.

The examination of the current advancements in causal inference and large language models serves a dual purpose. Firstly, it enhances our comprehension of how these two fields mutually benefit from each other. Secondly, it catalyzes the emergence of novel questions, propelling us closer to achieving Artificial General

Intelligence. Moreover, this exploration has the potential to extend into diverse realms and find applications in real-world scenarios, showcasing the far-reaching implications of the synergy between causal inference and LLM.

References

[1] Sara Abdali et al. "Extracting Self-Consistent Causal Insights from Users Feedback with LLMs and In-context Learning". In: *arXiv preprint arXiv:2312.06820* (2023).

[2] Josh Achiam et al. "GPT-4 Technical Report". In: *arXiv preprint arXiv:2303.08774* (2023).

[3] Nur Arifin Akbar et al. "Deep Learning of a Pre-trained Language Model's Joke Classifier Using GPT-2". In: *Journal of Hunan University Natural Sciences* 48.8 (2021).

[4] Alessandro Antonucci, Gregorio Piqu'e, and Marco Zaffalon. "Zero-shot Causal Graph Extrapolation from Text via LLMs". In: 2023. URL: https://api.semanticscholar.org/CorpusID:266521610.

[5] Vahan Arsenyan and Davit Shahnazaryan. "Large Language Models for Biomedical Causal Graph Construction". In: *arXiv preprint arXiv:2301.12473* (2023).

[6] Jinze Bai et al. "Qwen technical report". In: *arXiv preprint arXiv:2309.16609* (2023).

[7] Taiyu Ban et al. "Causal Structure Learning Supervised by Large Language Model". In: *arXiv preprint arXiv:2311.11689* (2023).

[8] Taiyu Ban et al. "From query tools to causal architects: Harnessing large language models for advanced causal discovery from data". In: *arXiv preprint arXiv:2306.16902* (2023).

[9] Rongzhou Bao, Jiayi Wang, and Hai Zhao. "Defending pre-trained language models from adversarial word substitutions without performance sacrifice". In: *arXiv preprint arXiv:2105.14553* (2021).

[10] Nora Belrose et al. "Eliciting latent predictions from transformers with the tuned lens". In: *arXiv preprint arXiv:2303.08112* (2023).

[11] Lorenzo Betti et al. "Relevance-based Infilling for Natural Language Counterfactuals". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2023, pp. 88–98.

[12] Amrita Bhattacharjee et al. "LLMs as Counterfactual Explanation Modules: Can ChatGPT Explain Black-box Text Classifiers?" In: *arXiv preprint arXiv:2309.13340* (2023).

[13] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[14] Sébastien Bubeck et al. "Sparks of artificial general intelligence: Early experiments with gpt-4". In: *arXiv preprint arXiv:2303.12712* (2023).

[15] Boxi Cao et al. "Can prompt probe pretrained language models? understanding the invisible risks from a causal view". In: *arXiv preprint arXiv:2203.12258* (2022).

[16] Nicolas Carion et al. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[17] Hang Chen et al. "Learning a Structural Causal Model for Intuition Reasoning in Conversation". In: *arXiv preprint arXiv:2305.17727* (2023).

[18] Yanda Chen et al. "Do models explain themselves? counterfactual simulatability of natural language explanations". In: *arXiv preprint arXiv:2307.08678* (2023).

[19] Zeming Chen et al. "DISCO: distilling counterfactuals with large language models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 5514–5528.

[20] Robert Dale. "GPT-3: What's it good for?" In: *Natural Language Engineering* 27.1 (2021), pp. 113–118.

[21] Ernest Davis and Gary Marcus. "Commonsense reasoning and commonsense knowledge in artificial intelligence". In: *Communications of the ACM* 58.9 (2015), pp. 92–103.

[22] Lei Ding et al. "Word embeddings via causal inference: Gender bias reducing and semantic information preserving". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 11. 2022, pp. 11864–11872.

[23] Amir Feder et al. "Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond". In: *Transactions of the Association for Computational Linguistics* 10 (2022). Ed. by Brian Roark and Ani Nenkova, pp. 1138–1158. DOI: 10.1162/tacl_a_00511. URL: https://aclanthology.org/2022.tacl-1.66.

[24] Amir Feder et al. "Causal-structure Driven Augmentations for Text OOD Generalization". In: *arXiv preprint arXiv:2310.12803* (2023).

[25] Eve Fleisig et al. "Fair-Prism: Evaluating fairness-related harms in text generation". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*. 2023.

[26] Luciano Floridi and Massimo Chiriatti. "GPT-3: Its nature, scope, limits, and consequences". In: *Minds and Machines* 30 (2020), pp. 681–694.

[27] Jinglong Gao et al. "Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation". In: *arXiv preprint arXiv:2305.07375* (2023).

[28] Yair Gat et al. "Faithful Explanations of Black-box NLP Models Using LLM-generated Counterfactuals". In: *arXiv preprint arXiv:2310.00603* (2023).

[29] Atticus Geiger et al. "Finding alignments between interpretable causal variables and distributed neural representations". In: *arXiv preprint arXiv:2303.02536* (2023).

[30] Tianrui Guan et al. "LOC-ZSON: Language-driven Object-Centric Zero-Shot Object Retrieval and Navigation". In: *arXiv preprint arXiv* (2024).

[31] Arnav Gudibande et al. "The false promise of imitating proprietary llms". In: *arXiv preprint arXiv:2305.15717* (2023).

[32] Riccardo Guidotti et al. "A survey of methods for explaining black box models". In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.

[33] Wes Gurnee et al. "Finding Neurons in a Haystack: Case Studies with Sparse Probing". In: *arXiv preprint arXiv:2305.01610* (2023).

[34] Jordan Hoffmann et al. "Training compute-optimal large language models". In: *arXiv preprint arXiv:2203.15556* (2022).

[35] Yifan Hou et al. "Towards a Mechanistic Interpretation of Multi-Step Reasoning Capabilities of Language Models". In: *arXiv preprint arXiv:2310.14491* (2023).

[36] Yinya Huang et al. "CLOMO: Counterfactual Logical Modification with Large Language Models". In: *arXiv preprint arXiv:2311.17438* (2023).

[37] David F Jenny et al. "Navigating the Ocean of Biases: Political Bias Attribution in Language Models via Causal Structures". In: *arXiv preprint arXiv:2311.08605* (2023).

[38] Zhenlan Ji et al. "Benchmarking and Explaining Large Language Model-based Code Generation: A Causality-Centric Approach". In: *arXiv preprint arXiv:2310.06680* (2023).

[39] Robin Jia et al. "Certified robustness to adversarial word substitutions". In: *arXiv preprint arXiv:1909.00986* (2019).

[40] Zhijing Jin et al. "Can Large Language Models Infer Causation from Correlation?" In: *arXiv preprint arXiv:2306.05836* (2023).

[41] Zhijing Jin et al. "Cladder: A benchmark to assess causal reasoning capabilities of language models". In: *arXiv preprint arXiv:2312.04350* (2023).

[42] Chandra Khatri et al. "Advancing the state of the art in open domain dialog systems through the alexa prize". In: *arXiv preprint arXiv:1812.10757* (2018).

[43] Yuheun Kim et al. "Can ChatGPT Understand Causal Language in Science Claims?" In: *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. 2023, pp. 379–389.

[44] Emre Kıcıman et al. "Causal reasoning and large language models: Opening a new frontier for causality". In: *arXiv preprint arXiv:2305.00050* (2023).

[45] Dohwan Ko et al. "Large Language Models are Temporal and Causal Reasoners for Video Question Answering". In: *arXiv preprint arXiv:2310.15747* (2023).

[46] Manabu Kuroki and Judea Pearl. "Measurement bias and effect restoration in causal inference". In: *Biometrika* 101.2 (2014), pp. 423–437.

[47] Jia Li and Xiang Li. "Relation-Oriented: Toward Knowledge-Aligned Causal AI". In: *arXiv preprint arXiv:2307.16387* (2023).

[48] Junnan Li et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *arXiv preprint arXiv:2301.12597* (2023).

[49] Xiaochuan Li et al. "Image Content Generation with Causal Reasoning". In: *arXiv preprint arXiv:2312.07132* (2023).

[50] Yongqi Li et al. "Large Language Models as Counterfactual Generator: Strengths and Weaknesses". In: *arXiv preprint arXiv:2305.14791* (2023).

[51] Zhang Li et al. "Monkey: Image resolution and text label are important things for large multi-modal models". In: *arXiv preprint arXiv:2311.06607* (2023).

[52] Zongxia Li et al. "Towards understanding in-context learning with contrastive demonstrations and saliency maps". In: *arXiv preprint arXiv:2307.05052* (2023).

[53] Percy Liang et al. "Holistic evaluation of language models". In: *arXiv preprint arXiv:2211.09110* (2022).

[54] Zachary C Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[55] Fuxiao Liu et al. "Aligning Large Multi-Modal Model with Robust Instruction Tuning". In: *arXiv preprint arXiv:2306.14565* (2023).

[56] Fuxiao Liu et al. "HallusionBench: You See What You Think? Or You Think What You See? An Image-Context Reasoning Benchmark Challenging for GPT-4V (ision), LLaVA-1.5, and Other Multi-modality Models". In: *arXiv preprint arXiv:2310.14566* (2023).

[57] Fuxiao Liu et al. "MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning". In: *arXiv preprint arXiv:2311.10774* (2023).

[58] Fuxiao Liu et al. "Visual news: Benchmark and challenges in news image captioning". In: *arXiv preprint arXiv:2010.03743* (2020).

[59] Haotian Liu et al. "Visual instruction tuning". In: *arXiv preprint arXiv:2304.08485* (2023).

[60] Xiao Liu et al. "The Magic of IF: Investigating Causal Reasoning Abilities in Large Language Models of Code". In: *arXiv preprint arXiv:2305.19213* (2023).

[61] Stephanie Long et al. "Can large language models build causal graphs?" In: *arXiv preprint arXiv:2303.05279* (2023).

[62] Yujie Lu et al. "Neuro-Symbolic Procedural Planning with Commonsense Prompting". In: *arXiv preprint arXiv:2206.02928* (2022).

[63] Muhammad Maaz et al. "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models". In: *arXiv preprint arXiv:2306.05424* (2023).

[64] Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. "What Makes Chain-of-Thought Prompting Effective? A Counterfactual Study". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pp. 1448–1535.

[65] Rahul Madhavan et al. "CFL: Causally Fair Language Models Through Token-level Attribute Controlled Generation". In: *arXiv preprint arXiv:2306.00374* (2023).

[66] James Manyika and Sissie Hsiao. "An overview of bard: an early experiment with generative ai". In: *AI. Google Static Documents* 2 (2023).

[67] Rui Mao et al. "GPTEval: A survey on assessments of ChatGPT and GPT-4". In: *arXiv preprint arXiv:2308.12488* (2023).

[68] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. "An empirical survey of the effectiveness of debiasing techniques for pre-trained language models". In: *arXiv preprint arXiv:2110.08527* (2021).

[69] Zhao Meng et al. "Self-Supervised Contrastive Learning with Adversarial Perturbations for Defending Word Substitution-based Attacks". In: *arXiv preprint arXiv:2107.07610* (2021).

[70] Xin Miao, Yongqi Li, and Tieyun Qian. "Generating Commonsense Counterfactuals for Stable Relation Extraction". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 5654–5668.

[71] Narmada Naik et al. "Applying Large Language Models for Causal Structure Learning in Non Small Cell Lung Cancer". In: *arXiv preprint arXiv:2311.07191* (2023).

[72] Allen Nie et al. "MoCa: Measuring Human-Language Model Alignment on Causal and Moral Judgment Tasks". In: *arXiv preprint arXiv:2310.19677* (2023).

[73] Liangming Pan et al. "Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning". In: *arXiv preprint arXiv:2305.12295* (2023).

[74] Nick Pawlowski et al. "Answering Causal Questions with Augmented LLMs". In: (2023).

[75] Judea Pearl. *Causality*. Cambridge university press, 2009.

[76] Judea Pearl. "Graphical models for probabilistic and causal reasoning". In: *Quantified representation of uncertainty and imprecision* (1998), pp. 367–389.

[77] Baolin Peng et al. "Instruction tuning with gpt-4". In: *arXiv preprint arXiv:2304.03277* (2023).

[78] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[79] Angelika Romanou et al. "CRAB: Assessing the Strength of Causal Relationships Between Real-world Events". In: *arXiv preprint arXiv:2311.04284* (2023).

[80] Donald B Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5 (1974), p. 688.

[81] Timo Schick and Hinrich Schütze. "Exploiting cloze questions for few shot text classification and natural language inference". In: *arXiv preprint arXiv:2001.07676* (2020).

[82] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. "Evaluating Gender Bias in Machine Translation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1679–1684. DOI: 10.18653/v1/P19-1164. URL: https://aclanthology.org/P19-116

[83] Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. *A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis*. 2023. arXiv: 2305.15054 [cs.CL].

[84] Shane Storks, Qiaozi Gao, and Joyce Y Chai. "Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches". In: *arXiv preprint arXiv:1904.01172* (2019), pp. 1–60.

[85] Hung-Ting Su et al. "Language Models are Causal Knowledge Extractors for Zero-shot Video Question Answering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 4950–4959.

[86] Yan Tai et al. "Link-context learning for multimodal llms". In: *arXiv preprint arXiv:2308.07891* (2023).

[87] Juanhe TJ Tan. "Causal Abstraction for Chain-of-Thought Reasoning in Arithmetic Word Problems". In: *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. 2023, pp. 155–168.

[88] Ziyi Tang et al. "Towards causalgpt: A multi-agent approach for faithful knowledge reasoning via promoting causal consistency in llms". In: *arXiv preprint arXiv:2308.11914* (2023).

[89] Rohan Taori et al. "Alpaca: A strong, replicable instruction-following model". In: *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html* 3.6 (2023), p. 7.

[90] Hugo Touvron et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[91] Ruibo Tu, Chao Ma, and Cheng Zhang. "Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis". In: *arXiv preprint arXiv:2301.13819* (2023).

[92] Aniket Vashishtha et al. "Causal Inference Using LLM-Guided Discovery". In: *arXiv preprint arXiv:2310.15117* (2023).

[93] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[94] Yuxuan Wan et al. "Biasasker: Measuring the bias in conversational ai system". In: *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 2023, pp. 515–527.

[95] Fei Wang et al. "A Causal View of Entity Bias in (Large) Language Models". In: *arXiv preprint arXiv:2305.14695* (2023).

[96] Xiyao Wang et al. "Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences". In: *arXiv preprint arXiv:2401.10529* (2024).

[97] Jason Wei et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 24824–24837.

[98] Moritz Willig et al. "Probing for correlations of causal facts: Large language models and causality". In: (2022).

[99] Sewall Wright. "The method of path coefficients". In: *The annals of mathematical statistics* 5.3 (1934), pp. 161–215.

[100] Xiyang Wu et al. "On the Safety Concerns of Deploying LLMs/VLMs in Robotics: Highlighting the Risks and Vulnerabilities". In: *arXiv preprint arXiv:2402.10340* (2024).

[101] Zhengxuan Wu et al. "Interpretability at scale: Identifying causal mechanisms in alpaca". In: *arXiv preprint arXiv:2305.08809* (2023).

[102] Yijia Xiao et al. "Large language models can be good privacy protection learners". In: *arXiv preprint arXiv:2310.02469* (2023).

[103] Yuxi Xie, Guanzhen Li, and Min-Yen Kan. "ECHo: Event Causality Inference via Human-centric Reasoning". In: *arXiv preprint arXiv:2305.14740* (2023).

[104] Jinghang Xu et al. "A review of dataset and labeling methods for causality extraction". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 1519–1531.

[105] Jie Yang, Soyeon Caren Han, and Josiah Poon. "A survey on extraction of causal relations from natural language text". In: *Knowledge and Information Systems* 64.5 (2022), pp. 1161–1186.

[106] Qinghao Ye et al. "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration". In: *arXiv preprint arXiv:2311.04257* (2023).

[107] Wenhao Yu et al. "IfQA: A Dataset for Open-domain Question Answering under Counterfactual Presuppositions". In: *arXiv preprint arXiv:2305.14010* (2023).

[108] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. "A survey on causal discovery: theory and practice". In: *International Journal of Approximate Reasoning* 151 (2022), pp. 101–129.

[109] Matej Zečević et al. "Causal parrots: Large language models may talk causality but are not causal". In: *arXiv preprint arXiv:2308.13067* (2023).

[110] Jingying Zeng and Run Wang. “A survey of causal inference frameworks”. In: *arXiv preprint arXiv:2209.00869* (2022).

[111] Wei Zeng et al. “Pangua: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation”. In: *arXiv preprint arXiv:2104.12369* (2021).

[112] Runtian Zhai et al. “Macer: Attack-free and scalable robust training via maximizing certified radius”. In: *arXiv preprint arXiv:2001.02378* (2020).

[113] Jiaqi Zhang et al. “Towards Causal Foundation Model: on Duality between Causal Inference and Attention”. In: *arXiv preprint arXiv:2310.00809* (2023).

[114] Jiayao Zhang et al. “Rock: Causal inference principles for reasoning about commonsense causality”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 26750–26771.

[115] Li Zhang et al. “Causal Reasoning of Entities and Events in Procedural Texts”. In: *arXiv preprint arXiv:2301.10896* (2023).

[116] Shuo Zhang et al. “Mitigating Language Model Hallucination with Interactive Question-Knowledge Alignment”. In: *arXiv preprint arXiv:2305.13669* (2023).

[117] Yanzhe Zhang et al. “Llavar: Enhanced visual instruction tuning for text-rich image understanding”. In: *arXiv preprint arXiv:2306.17107* (2023).

[118] Haiteng Zhao et al. “Certified robustness against natural language attacks by causal intervention”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 26958–26970.

[119] Haiyan Zhao et al. “Explainability for large language models: A survey”. In: *ACM Transactions on Intelligent Systems and Technology* (2023).

[120] Qinlin Zhao et al. “Competeai: Understanding the competition behaviors in large language model-based agents”. In: *arXiv preprint arXiv:2310.17512* (2023).

[121] Shitian Zhao et al. “Causal-CoG: A Causal-Effect Look at Context Generation for Boosting Multi-modal Language Models”. In: *arXiv preprint arXiv:2312.06685* (2023).

[122] Wei Zhao, Zhe Li, and Jun Sun. “Causality Analysis for Evaluating the Security of Large Language Models”. In: *arXiv preprint arXiv:2312.07876* (2023).

[123] Junhao Zheng et al. “Preserving Commonsense Knowledge from Pre-trained Language Models via Causal Inference”. In: *arXiv preprint arXiv:2306.10790* (2023).

[124] Fan Zhou et al. “Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 4227–4241.

[125] Yuhang Zhou, Suraj Maharjan, and Beiye Liu. “Scalable prompt generation for semi-supervised learning with language models”. In: *arXiv preprint arXiv:2302.09236* (2023).

[126] Yuhang Zhou et al. “Explore Spurious Correlations at the Concept Level in Language Models for Text Classification”. In: *arXiv preprint arXiv:2311.08648* (2023).

[127] Deyao Zhu et al. “Minigpt-4: Enhancing vision-language understanding with advanced large language models”. In: *arXiv preprint arXiv:2304.10592* (2023).