Large Language Models for Causal Discovery: Current Landscape and Future Directions

Guangya Wan[1] , **Yunsheng Lu**[2] , **Yuqi Wu**[3] , **Mengxuan Hu**[1] , **Sheng Li**[1]
[1]School of Data Science, University of Virginia
[2]Department of Statistics, University of Chicago
[3]Department of Electrical and Computer Engineering, University of Alberta
{wxr9et, qtq7su, shengli}@virgnia.edu, yunslu@uchicago.edu, yuqi14@ualberta.ca

Abstract

Causal discovery (CD) and Large Language Models (LLMs) have emerged as transformative fields in artificial intelligence that have evolved largely independently. While CD specializes in uncovering cause-effect relationships from data, and LLMs excel at natural language processing and generation, their integration presents unique opportunities for advancing causal understanding. This survey examines how LLMs are transforming CD across three key dimensions: direct causal extraction from text, integration of domain knowledge into statistical methods, and refinement of causal structures. We systematically analyze approaches that leverage LLMs for CD tasks, highlighting their innovative use of metadata and natural language for causal inference. Our analysis reveals both LLMs' potential to enhance traditional CD methods and their current limitations as imperfect expert systems. We identify key research gaps, outline evaluation frameworks and benchmarks for LLM-based causal discovery, and advocate future research efforts for leveraging LLMs in causality research. As the first comprehensive examination of the synergy between LLMs and CD, this work lays the groundwork for future advances in the field.

1 Introduction

Uncovering causal relationships—understanding why things happen—is fundamental to scientific discovery and informed decision-making across diverse domains. From discovering the causes of diseases and developing effective treatments to optimizing complex systems like city traffic flow or global supply chains, knowing why something occurs is crucial for effective intervention [Kuang *et al.*, 2020]. Causal Discovery (CD) has long relied on two pillars: statistical methods for data analysis and domain experts for knowledge integration [Pearl, 2009]. While domain experts provide invaluable insights drawn from years of experience and deep understanding, their involvement often creates bottlenecks in the discovery process. Consulting experts is time-consuming, expensive, and inherently limited by human availability and potential biases. Meanwhile, Statistical CD methods (SCD) [Shimizu *et al.*, 2011; Scutari and Denis, 2014; Zheng *et al.*, 2018], while mathematically rigorous, often fall short in real-world scenarios. They typically demand vast amounts of high-quality data, which is often unavailable or expensive to acquire. Furthermore, they struggle to disentangle complex temporal dynamics inherent in many real-world systems, where causes and effects unfold over time and influence each other in intricate ways [Ban *et al.*, 2023a]. Specifically, these methods frequently produce multiple, equally plausible causal explanations, traditionally requiring expert intervention to resolve these ambiguities.

Large Language Models (LLMs) offer a transformative approach to the challenges of causal discovery, potentially serving as valuable tools to complement and augment human expertise. Their ability to process and synthesize massive amounts of text—effectively distilling knowledge from countless documents, research papers, and expert opinions—makes them powerful aids for enhancing expert-level reasoning [Petroni *et al.*, 2019]. Unlike traditional domain experts with specialized knowledge in specific areas, LLMs can provide broad perspective across multiple domains concurrently, working alongside human experts rather than replacing them. They can integrate information from diverse sources to help identify potential causal relationships that might benefit from further expert validation. For example, when analyzing urban traffic, an LLM could rapidly synthesize insights from traffic engineering papers, weather studies, and city planning documents to suggest potential causal factors for human experts to accelerate what could otherwise take weeks of literature review.

The integration of LLMs into CD represents a paradigm shift from both purely statistical approaches and traditional expert-dependent methods, manifesting in three primary ways: (1) LLMs can directly infer causal graphs or subgraph structures from natural language descriptions and domain knowledge [Jin *et al.*, 2023b], effectively automating the initial expert hypothesis generation phase. (2) LLMs can function as posterior correction mechanisms, validating and refining causal relationships identified by SCD methods against their extensive knowledge base [Long *et al.*, 2024], similar to how experts would review and adjust statistical findings. (3) They can serve as comprehensive prior information sources for traditional SCD algorithms, providing domain knowledge and contextual constraints before statisti-

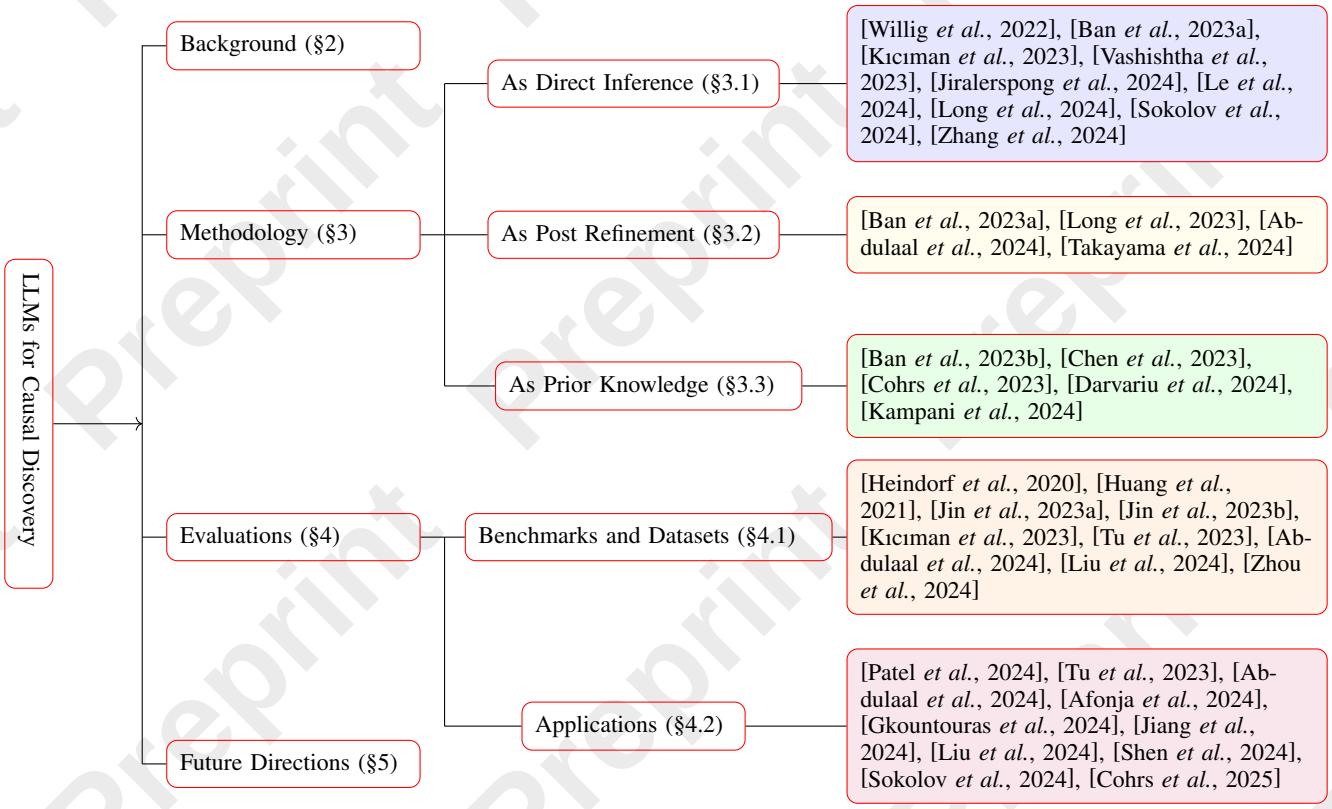

Figure 1: Overview of LLM-based causal discovery methods, categorized by their role: direct evaluation, prior knowledge augmentation, and post-hoc refinement. Evaluation strategies and future directions are also outlined.

cal analysis [Takayama *et al.*, 2024]. This systematic enhancement of human expert involvement with LLM-based assistance not only accelerates the discovery process but also makes sophisticated causal analysis accessible to a broader range of researchers and without domain expertise.

Several surveys have examined how large language models (LLMs) relate to causality. However, there is still a lack of detailed survey specifically on *causal discovery* in the era of LLMs. Traditional causal discovery surveys like [Glymour *et al.*, 2019] extensively explore connections with machine learning and deep learning approaches but lacks discussion with the emergence of LLMs in this domain. [Zhao *et al.*, 2023] pioneered the discussion of LLMs in causal reasoning but primarily focuses on broader tasks such as counterfactual reasoning, cause attribution, and causal effect estimation. More recent work by [Zhang *et al.*, 2024] examines causality in LLMs but provides a limited analysis of how these models fundamentally transform causal discovery methods. Similarly, [Yu *et al.*, 2024] offers a comprehensive review of improving LLMs' causal reasoning capabilities but lacks specific focus on causal discovery tasks and their integration with traditional causal theory. To address this gap in literature, our survey examines the emerging intersection of LLMs and CD, providing a systematic framework for understanding their integration. We begin in Section 2 with background on both LLMs and CD, bridging knowledge gaps for researchers from either field. Section 3 analyzes how LLMs can enhance CD through direct inference, prior knowledge integration, and structural refinement, while acknowledging the methodological challenges each approach faces. Section 4 evaluates benchmark datasets and showcases applications across diverse domains, from healthcare to social sciences. Section 5 examines current limitations, explores open questions about LLMs' causal reasoning capabilities, and identifies promising research directions that could advance this rapidly evolving field.

2 Background

This section lays the groundwork for understanding fundamental concepts and methodologies of causal inference and discovery, including essential notations and definitions in causal discovery to prepare readers better understand and subsequent technical sections.

2.1 Graphical Models and Structure Learning

Directed Acyclic Graphs (DAGs) provide the foundational framework for causal modeling, defined as ordered pairs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents variables and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ contains directed edges without cycles. Structure learning aims to estimate these graphs by capturing essential dependence relationships in data [Drton and Maathuis, 2017]. While any probability measure $\mathbb{P}(\mathbf{X})$ can be factored as $\mathbb{P}(\mathbf{X}) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1)\cdots\mathbb{P}(X_n|X_1,\ldots,X_{n-1})$, this

$$X_1 \rightarrow X_2 \rightarrow X_3 \qquad X_1 \leftarrow X_2 \leftarrow X_3$$

Figure 2: Two Markov equivalent DAGs

complete representation introduces unnecessary dependencies. The goal is identifying minimal structures where edges $X_i \rightarrow X_j$ represent only necessary conditional dependencies [Pearl, 2009]. A fundamental limitation is that different DAGs can encode identical conditional independence relations, forming Markov equivalence classes. For example, given variables with relations $X_1 \not\perp\!\!\!\perp X_2, X_2 \not\perp\!\!\!\perp X_3, X_1 \not\perp\!\!\!\perp X_3, X_1 \perp\!\!\!\perp X_3 \mid X_2$, multiple network structures (Figure 2) can represent these relationships, showing that causal direction cannot be uniquely determined from data alone.

2.2 Causal Discovery: Methods and Evaluation

A causal graph is a DAG where edges represent direct causal relationships—an edge $X_i \rightarrow X_j$ indicates that X_i is a direct cause of X_j. **Causal Discovery (CD)** is the systematic process of uncovering these causal relationships from observational data [Glymour *et al.*, 2019], which typically involves learning Bayesian Network structures with causal interpretations when all common causes are observed. These discovered structures serve as foundations for downstream applications including effect inference [Kuang *et al.*, 2020] and prediction [Chu *et al.*, 2023]. Formally, a **Structural Causal Model (SCM)** $M = (V, U, F, P)$ provides the framework for representing these systems [Pearl, 2009], where V represents observable variables, U unobservable variables, F causal mechanisms, and P the probability distribution. Each variable $X_i \in V$ is determined by a function $f_i \in F$ of its direct causes $Pa(X_i)$ and an independent noise term $U_i \in U$: $X_i = f_i(Pa(X_i), U_i)$. The goal of causal discovery is to recover these structural components, particularly the parent sets defining the underlying graph structure.

Statistical approaches to causal discovery have evolved into two methodologies: **Constraint-based methods** like the PC Algorithm [Spirtes *et al.*, 2001] that test conditional independence relationships ($X \perp Y|Z$) to remove edges and orient directions, and **Score-based methods** that optimize scoring functions balancing model fit against complexity, exemplified by NOTEARS [Zheng *et al.*, 2018]. Causal discovery with LLMs addresses two distinct tasks with different evaluation requirements: **Causal Order Predictions** (pairwise discovery) evaluates direct relationships between variable pairs using standard classification metrics like accuracy and F1-score, while **Full Graph Discovery** constructs complete causal networks using Structural Hamming Distance (SHD) and Normalized Hamming Distance (NHD) [Tsamardinos *et al.*, 2006], which measure edge operations needed to transform a learned graph into the true causal structure.

3 LLMs for Causal Discovery

Large Language Models (LLMs) have revolutionized natural language processing through their advanced transformer architecture, demonstrating remarkable capabilities in reasoning, knowledge acquisition, and cross-domain generalization [Zhao *et al.*, 2023], while serving as reliable knowledge bases [Petroni *et al.*, 2019] suitable for causal discovery tasks, enabling them to assume the role of human domain experts as illustrated in Figure 3. We discuss three primary approaches to incorporating LLMs in causal discovery: (a) direct inference, (b) posterior refinement on derived causal structures, and (c) knowledge integration as prior for generating causal structures, with Table 1 providing practical examples of prompts used in these approaches.

3.1 LLMs as Direct Inference

Direct causal discovery leverages LLMs' extensive knowledge acquired during pre-training to serve as automated domain experts capable of reasoning about causal relationships. Unlike traditional methods that rely on statistical patterns or require extensive human expert consultation, LLMs can utilize their broad understanding of domain concepts, scientific principles, and real-world relationships to infer causality at scale. The fundamental setting involves providing LLMs with meta-data such as the descriptive texts $T = \{t_1, t_2, \ldots, t_n\}$ for variables $X = \{x_1, x_2, \ldots, x_n\}$. By comprehending these descriptions and applying learned knowledge, LLMs identify causal statements denoted as $S = \{(x_i, x_j)\}$, where (x_i, x_j) indicates that x_i causes x_j. This approach effectively transforms LLMs into scalable meta-data experts who can reason about causality with both breadth and precision.

Two primary approaches have emerged in this direction. First, **causal order prediction**, pioneered by [Willig *et al.*, 2022] and advanced by [Kıcıman *et al.*, 2023], focuses on determining pairwise causal relationships through direct LLM queries, primaliry using chain-of-thought style prompting [Wei *et al.*, 2022]. Second, **complete and partial causal graph discovery** methods aim to identify broader causal structures through iterative pairwise discovery [Long *et al.*, 2024; Kıcıman *et al.*, 2023], though this naturally introduces computational challenges scaling quadratically with the number of variables. To address these efficiency barriers, [Jiralerspong *et al.*, 2024] reduced computational complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$ through a structured three-phase process: root cause identification, relationship expansion, and logical consistency verification. For high-dimensional settings, [Sokolov *et al.*, 2024] developed a scalable solution using hierarchical clustering based on semantic similarity, efficiently discovering causal relationships by first analyzing within-cluster connections before determining inter-cluster causality. To enhance the reliability of these approaches, researchers have pursued several complementary directions: systematic verification to mitigate hallucination [Ji *et al.*, 2023], specialized fine-tuning for causal reasoning [Le *et al.*, 2024], and structured prompting frameworks for consistent causal extraction [Zhang *et al.*, 2024; Vashishtha *et al.*, 2023].

3.2 LLMs as Posterior Correction

LLMs can serve as a expert judge by correcting and refining the learned causal structures from the traditional statistical causal discovery (SCD) methods based on contextual reasoning or additional data. Recall that most of constraint-based and score-based methods can only identify a BN up to its Markov equivalence class. Given the set of all conditional independence relations, [Long *et al.*, 2023] introduces a method

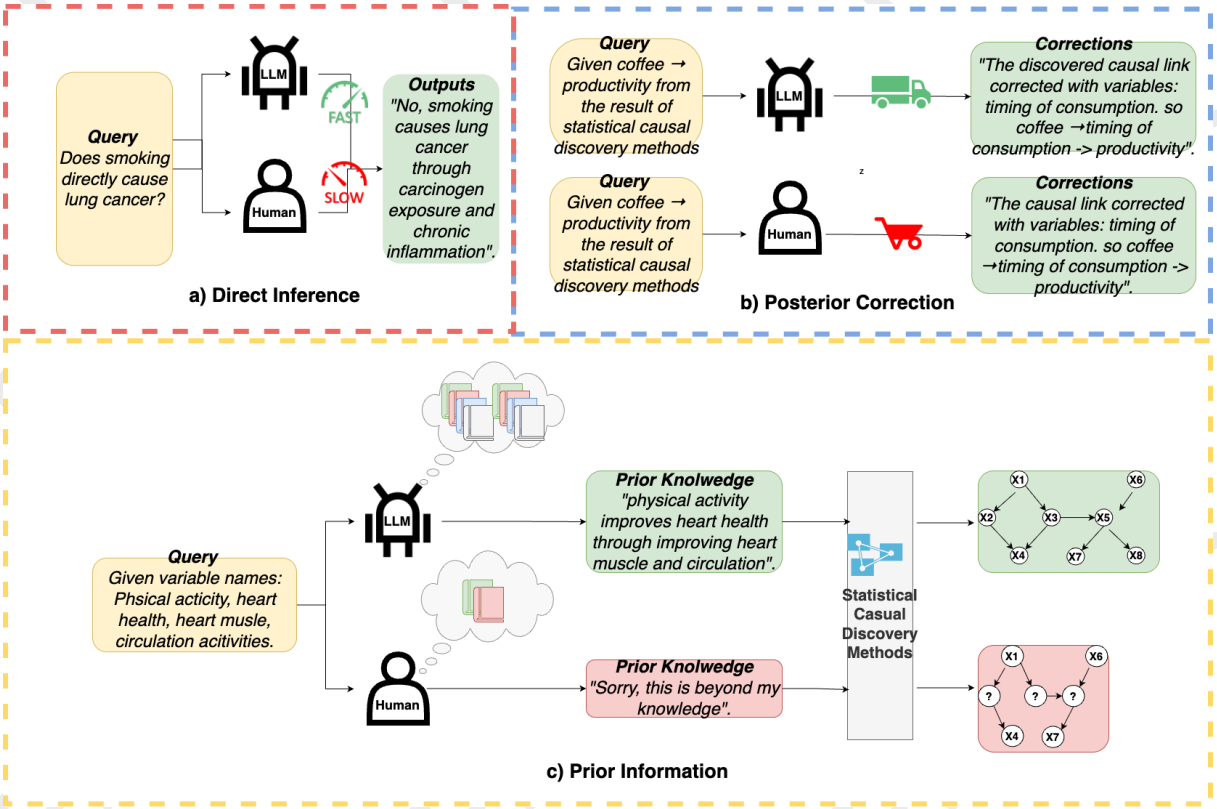

Figure 3: Three distinct approaches for applying LLMs in causal discovery: (a) Direct causal inference without observational data, (b) Post refinement of statistically derived causal structures, and (c) Integration of prior knowledge into traditional statistical methods. The figure highlights the increasing automation and precision of causal discovery through LLMs, reducing the need for manual expert input.

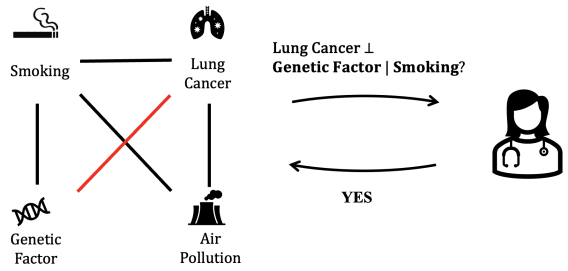

Figure 4: General Workflow of Applying LLM on Conditional Independence Tests to Solve Constraint-Based Discovery Tasks.

that uses LLM as an imperfect expert to progressively reduce the number of possible causal structures within the Markov equivalence class, while controlling the risk of misorienting edges. However, its application to larger datasets remains unproven and is likely hindered by the approach's complexity and computational demands. To refine causal structures beyond the Markov equivalence class using LLMs, [Long *et al.*, 2023] formulated a constrained discrete optimization problem: $\min_S |\mathcal{M}^{E,S}|$ subject to $\mathbb{P}(G^\star \in \mathcal{M}^{E,S}) \geq 1 - \eta$, where $\mathcal{M}^{E,S}$ denotes the refined equivalence class after incorporating expert-guided orientations, and η controls the risk of excluding the true graph $G^\star$. Define $\mathcal{U}$ to be all undirected edges whose orientations vary across DAGs in the MEC, two greedy strategies are proposed: one prioritizes reducing the size of the equivalence class, and the other minimizes the risk of eliminating, both following the iterative structure 1:

Instead of deriving the true causal graph beyond its Markov equivalence class, the Iterative LLM Supervised CSL Framework (ILS-CSL) by [Ban *et al.*, 2023a] refines a partially learned DAG by leveraging LLM feedback iteratively to correct edge orientations. This approach efficiently integrates expert knowledge while avoiding exhaustive pairwise queries, ensuring a more accurate and robust causal structure without requiring full causal discovery from scratch.

[Takayama *et al.*, 2024] introduced a framework in which large language models (LLMs) function both as post hoc refiners and prior knowledge generators, described in Algorithm 2: Initially, the raw adjacency matrix $\hat{G}_0$, derived from certain SCD methods without prior knowledge, is input into the LLMs. For each potential edge, the LLM is queried multiple times using pairwise prompts, with each response being a binary decision (i.e., "yes" or "no"). The probability matrix P is then constructed by aggregating these responses, and is subsequently converted into a deterministic prior knowledge matrix $\hat{G}$ based on predefined thresholds. Finally, $\hat{G}$ is incorporated into the original SCD method to infer the final causal graph. While this framework is primarily designed for LiNGAM, it is applicable to any causal discovery algorithm

Tasks	Prompt
Pairwise Discovery	"Which is more likely to be true: (A) lung cancer causes cigarette smoking, or (B) cigarette smoking causes lung cancer?"
Conditional Independence Set Test	As an expert in a specific field, you're asked to assess the statistical independence between two variables, potentially conditioned on another variable set. Your response, based on theoretical knowledge, should be a binary guess (YES or NO) and the probability of its correctness, formatted as: [ANSWER (PROBABILITY%)]. For example, [YES (70%)] or [NO (30%)].
Causal Validation	Given a statistical correlation between variables A and B, and their relationship with other variables in the system, determine if we can validly conclude that A causes B. Please provide your reasoning step by step and conclude with either 'Valid' or 'Invalid' for this causal inference.
Full Graph Discovery	As a domain expert, analyze cause-and-effect relationships among variables with given abbreviations and values. Interpret each variable and present the causal relationships as a directed graph, using edges to denote direct causality, e.g., $x_{i1} \to x_{j1}, \ldots, x_{im} \to x_{jm}$.

Table 1: A practical example of prompts with respect to various LLM causal discovery frameworks.

that relies on a matrix of pairwise edge scores.

Algorithm 1: MEC Refine with Imperfect Expert

Input: Initial MEC $\mathcal{M}$, expert E, tolerance η, strategy $S \in \{\texttt{size}, \texttt{risk}\}$
Output: Refined equivalence class $\mathcal{M}^{E,S}$
while $\mathbb{P}(G^\star \in \mathcal{M}) \geq 1 - \eta$ **do**
  foreach $p \in \mathcal{U}$ **do**
    Query orientation $E(p)$;
    Let $\mathcal{M}_p \leftarrow$ MEC after orienting p via $E(p)$ and applying Meek rules ;
    if $S = \texttt{size}$ **then**
      Score$[p] \leftarrow |\mathcal{M}_p|$
    else if $S = \texttt{risk}$ **then**
      Score$[p] \leftarrow 1 - \mathbb{P}(G^\star \in \mathcal{M}_p)$
  Select $p^\star = \arg\min_p$ Score$[p]$;
  Update $\mathcal{M} \leftarrow \mathcal{M}_{p^\star}$;
return $\mathcal{M}^{E,S} \leftarrow \mathcal{M}$

[Abdulaal *et al.*, 2024] introduced the Causal Modeling Agent (CMA), a novel framework for causal discovery that combines the metadata-based reasoning of LLMs with the data-driven power of Deep Structural Causal Models (DSCMs). CMA employs an LLM to propose an initial causal graph, which then informs the fitting of a DSCM to the data. The framework iteratively refines this graph in global and local phases, again using the LLM for both prior knowledge and a critic of the model's output, enabling the discovery of causal relationships in complex, multi-modal data. Unlike purely constraint-based or scoring-based methods, CMA integrates aspects of both while leveraging LLMs. Furthermore, it can generate chain graphs to account for unmeasured confounding and has demonstrated state-of-the-art performance on datasets like Arctic Sea [Huang *et al.*, 2021] and on synthetic data designed to prevent data leakage.

3.3 LLMs as Prior Knowledge

Similarly, LLMs can also be used in conjunction with traditional methods to provide source of prior knowledge by leveraging meta-data extracted from textual descriptions and domain-specific information. In [Ban *et al.*, 2023b], a set of variables **X** along with their descriptive texts **T** are provided as input to an LLM, which performs causal discovery by identifying direct relationships after comprehending the semantic meaning of each variable. Traditionally, the incorporation of prior knowledge into CD procedures follows two primary approaches: the hard constraint method and the soft constraint method. The hard constraint approach strictly enforces prior knowledge by eliminating edges in BNs that conflict with constraints derived from traditional causal discovery algorithms. However, this method lacks flexibility, as any spurious prior assumptions cannot be corrected during the learning process, potentially leading to erroneous causal structures. [Chen *et al.*, 2023] provides a systematic approach for detecting and correcting potentially erroneous prior knowledge derived from LLMs, thereby enhancing the reliability of utilizing LLMs for hard prior constraints. To address this, [Chen *et al.*, 2023] focus on a particularly impactful class of prior errors, termed *order-reversed priors*, where a prior $(X_j \to X_i)$ contradicts the true edge $(X_i \to X_j)$ in the underlying DAG G_0. The authors show that incorporating such priors under hard constraints induces a unique acyclic structure in the learned graph G_1, called a *quasi-circle*: a pair of distinct directed paths ℓ_1 and ℓ_2 from X_s to X_t sharing only their endpoints, with an example shown in Figure 5. Building on this insight, the authors propose a post-hoc correction strategy that iteratively identifies priors involved in quasi-circles, and replaces or removes them.

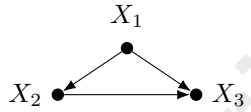

Figure 5: A minimal quasi-circle

In contrast, the soft constraint approach, though implemented through varying methodologies, aims to integrate prior knowledge in a fault-tolerant manner. [Ban *et al.*, 2023b] integrates LLM-derived prior into scoring functions, such as BDeu or BIC, as a regularization term, allowing flexibility in cases where LLM priors may be inconsistent with ob-

served data. Alternatively, [Darvariu *et al.*, 2024] introduces a probabilistic prior framework described in 2, where the LLM-derived priors consist of probabilities of the existence and direction of each edge in the causal graph, which are then incorporated into some traditional CD algorithms. In particular, the proposed approach is compatible with any causal discovery algorithm that relies on a pairwise edge score matrix, including LiNGAM and NOTEARS. Given the demonstrated effectiveness of integrating LLMs into differentiable causal discovery algorithms within this probabilistic framework, recent work by [Kampani *et al.*, 2024] further extends this paradigm by using LLMs to initialize the continuous optimization process. Unlike many previous methods that assume a parametric model (e.g., linear-Gaussian), focus primarily on continuous data, and compute likelihoods accordingly, the LLM-DCD framework targets discrete data using a non-parametric loss called `MLE-INTERP`. Notably, LLMs provide an initial estimate of the adjacency matrix A, thereby guiding the continuous optimization from an informed starting point. Although their results appears to be promising on discrete data, it's unclear whether a similar framework could be generalized to continuous data.

A novel application of LLMs as providers of prior information lies in conditional independence testing, a cornerstone of constraint-based causal discovery. Rather than using traditional statistical tests, LLMs can be queried with natural language prompts representing conditional independence relationships, effectively serving as an oracle. As shown in Figure 4, this enables constainted based algorithms like PC algorithm [Spirtes *et al.*, 2001] to leverage LLMs for guidance in causal graph construction. The chatPC method [Cohrs *et al.*, 2023] exemplifies this approach, integrating LLMs with the PC algorithm by transforming conditional independence tests (e.g., "Is X independent of Y given Z?") into natural language prompts. The LLM's responses then guide the PC algorithm's edge removal process. This work evaluates LLM performance on such queries, proposes a statistical aggregation method to combine multiple LLM responses for increased robustness, and analyzes the resulting causal graphs. Research indicates that LLMs tend to be more conservative in their independence judgments than human experts, yet still demonstrate evidence of causal reasoning.

Remark. Probabilistic prior knowledge has never been a strong focus in the causal discovery literature. Before the era of LLMs, expert priors were often assumed to be deterministically correct. Although recent efforts have introduced imperfect expert frameworks, they often rely on simplified assumptions. For instance, [Long *et al.*, 2023] models expert fallibility using a fixed noise level $\epsilon \in \{0.1, 0.3\}$, assuming a uniform error rate across all edge orientations. However, in practice, LLM accuracy is highly sensitive to prompt design, semantic ambiguity, and domain-specific difficulty—none of which are captured by a static ϵ. Meanwhile, [Chen *et al.*, 2023] focuses on quasi-circle detection to identify erroneous priors but limits the scope to cycles of length three, due to the exponential complexity $\mathcal{O}(n^{L-2})$ for general detection. As a result, when faced with novel tasks or unseen data, the reliability of LLM-derived causal knowledge remains uncertain and lacks any principled quantification.

Algorithm 2: Probabilistic LLM-driven Priors

Input: Data $\mathcal{D}$, metadata $\{\mu_i\}_{i=1}^{d}$, LLM expert E, score function f, budget B, prior strength τ
Output: Estimated DAG $\mathcal{G}^\star$
Initialize $\mathcal{G}^{(0)} \leftarrow (V, \emptyset)$, $P \in \mathbb{R}^{d\times d}$;
foreach *unordered pair* $(i, j), i \neq j$ **do**
  Query E on (μ_i, μ_j) to get $P_{i\to j}$;
  Set $P[i, j] \leftarrow P_{i\to j}$;
$\mathcal{G}^\star \leftarrow \mathcal{G}^{(0)}$, $s^\star \leftarrow \infty$;
for $b = 1$ **to** B **do**
  $\mathcal{G} \leftarrow \mathcal{G}^{(0)}$;
  while *termination not met* **do**
    Let $A \leftarrow$ valid acyclic edges not in $\mathcal{G}$;
    Sample $e_{i\to j} \sim \text{softmax}_{(i,j)\in A}(P[i, j]/\tau)$;
    $\mathcal{G} \leftarrow \mathcal{G} \cup \{e_{i\to j}\}$;
  Compute $s \leftarrow f(\mathcal{G})$;
  if $s < s^\star$ **then**
    $\mathcal{G}^\star \leftarrow \mathcal{G}$, $s^\star \leftarrow s$;
return $\mathcal{G}^\star$

4 Evaluations and Applications

4.1 Benchmarks and Datasets

Table 2 provides a comprehensive overview of benchmark datasets used to evaluate LLMs' causal reasoning capabilities. For each dataset, the table indicates (1) if for determining causal relationships between variable pairs, (2) full graph reconstruction, and (3) novel reasoning scenarios (testing on previously unseen causal patterns). The table also details key characteristics including the average number of nodes and edges per graph, along with the total number of graphs in the collection. Multi-graph benchmark datasets, such as CausalBench [Zhou *et al.*, 2024], are particularly noteworthy as they incorporate established causal networks dataset commonly tested in the literature, such as Asia [Pearl, 1988] and Insurance [Binder *et al.*, 1997], offering evaluation across diverse graph sizes and domains. Among these benchmarks, CORR2CAUSE [Jin *et al.*, 2023a] addresses a crucial aspect of causal reasoning: the ability to differentiate causation from correlation and can be further used for fine-tuning LLMs to enhance their causal inference capabilities from identifying purely correlational statements.

4.2 Applications

Causal discovery has been used as a crucial tool across numerous real-world domains, with LLM-based methods significantly expanding its capabilities and applications. For instance, [Gkountouras *et al.*, 2024] introduced a "causal world model" framework, connecting causal variables to natural language to improve reasoning in complex environments. To address the challenge of ill-defined high-level variables often found in real-world observational data, [Liu *et al.*, 2024] enables LLMs to propose such variables, effectively extending

Dataset Name	Work	Pair	Full	Novel	Domain	Avg. Nodes	Avg. Edges	Num. Graphs
Asia	[Pearl, 1988]	✓	✓	✓	Medical	8	8	–
Insurance	[Binder *et al.*, 1997]	✓	✓	✓	Business	27	52	–
CauseNet	[Heindorf *et al.*, 2020]	✓	×	×	Web/Mixed	12.2M	11.6M	–
Arctic Sea	[Huang *et al.*, 2021]	✓	✓	×	Climate	12	42	–
Neuropathic	[Tu *et al.*, 2023]	×	✓	×	Medical	222	770	–
Sangiovese	[Kıcıman *et al.*, 2023]	×	✓	×	Agriculture	15	55	–
Tübingen	[Kıcıman *et al.*, 2023]	✓	×	✓	Mixed Science	222	770	–
Alzheimer	[Abdulaal *et al.*, 2024]	✓	✓	✓	Medical	11	19	–
Multi-Graph Benchmark Datasets								
CLADDER	[Jin *et al.*, 2023a]	✓	✓	✓	Mixed	3.52	3.38	10,112
CORR2CAUSE	[Jin *et al.*, 2023b]	✓	✓	✓	Mixed	2-6	8.60	207,972
AppleGastronome	[Liu *et al.*, 2024]	✓	✓	✓	Food	6	5	200
CausalBench	[Zhou *et al.*, 2024]	✓	✓	✓	Mixed	2-109	11.6M	15

Note: Pair = Pairwise Discovery; Full = Full Graph Discovery; Novel = Involves a stimulator to regenerate data to avoid data leakage

Table 2: Summary of Benchmark Datasets for Evaluating Causal Discovery Tasks

causal discovery to unstructured data. These and other advances have facilitated LLM-guided causal discovery in fields like medicine [Tu *et al.*, 2023; Cohrs *et al.*, 2025], finance [Sokolov *et al.*, 2024], genetics [Afonja *et al.*, 2024], and health informatics [Patel *et al.*, 2024]. Further research explores LLM-driven causal discovery in multi-agent systems [Abdulaal *et al.*, 2024; Jiang *et al.*, 2024] and multi-modal data integration [Shen *et al.*, 2024], leveraging the richness of multi-modal data to provide additional information to better capture the complexity of real-world systems.

5 Challenges and Visions

Unified Evaluation and Domain-Specific Applications. A significant challenge in LLM-enhanced causal discovery is the lack of standardized evaluation protocols, making it difficult to establish true state-of-the-art performance. Researchers should utilize both synthetic datasets (avoiding data leakage) and established benchmarks while developing frameworks incorporating multiple performance metrics beyond accuracy. Simultaneously, we envision significant potential in domain-specialized models that better capture field-specific causal relationships in critical domains like healthcare and economics. These systems could be enhanced through integration with field-specific knowledge bases, specialized reasoning modules, and domain-specific RAG incorporating scientific literature—for instance, accessing pathway databases in biology or physics-based models in climate science. Continuous learning mechanisms could ensure these systems remain current with emerging research findings.

LLM for SCM Diagnosis. Future research should expand LLMs' role beyond identifying causal relationships to verifying underlying properties of Structural Causal Models. While current approaches primarily detect causal links [Jin *et al.*, 2023b], LLMs could verify crucial aspects like the nature of relationships (linear vs. nonlinear), functional forms, and noise distributions. This verification is particularly important since traditional methods depend on specific assumptions—DirectLiNGAM requires linear relationships [Shimizu *et al.*, 2011], and BIC scoring becomes less reliable with nonlinear effects [Peters *et al.*, 2017]. LLMs could leverage their language understanding to interpret domain knowledge about expected relationship characteristics, helping select appropriate algorithms and validate assumptions, thus improving reliability across diverse scenarios.

Explanability and Interpretbility. The capability of LLMs to perform genuine causal reasoning remains an open question, and studies [Zečević *et al.*, 2023; Feng *et al.*, 2024] have suggested LLMs may function more as pattern-matching systems reciting embedded knowledge rather than understanding true causality, functioning as 'causal parrots' without deeper understanding [Jin *et al.*, 2023a]. Empirical studies show LLMs often rely on correlational heuristics when faced with questions requiring understanding of confounding variables—indicating a gap between statistical association and causal understanding [Kıcıman *et al.*, 2023; Long *et al.*, 2024]. Future research should focus on: (1) developing interpretability methods to analyze how LLMs process causal relationships; (2) investigating relationships between pre-training data and causal capabilities; and (3) creating frameworks distinguishing between genuine reasoning and memorized patterns.

6 Conclusion

The integration of LLMs with causal discovery represents a promising yet challenging advancement in artificial intelligence. This survey has explored how LLMs can enhance traditional causal discovery through direct inference, knowledge integration, and structural refinement. However, as systems trained primarily on correlational data, LLMs face inherent limitations in genuine causal reasoning that requires interventions and counterfactuals. They may reproduce existing biases and generate plausible but incorrect causal relationships, highlighting critical areas for future research in understanding and improving LLMs' causal reasoning capabilities. Moving forward, we envision LLMs as complementary tools that assist human experts rather than replace them, ultimately accelerating scientific discovery while maintaining human expertise at the center of the causal discovery process.

Acknowledgments

The work is supported in part by the U.S. Office of Naval Research Award under Grant Number N00014-24-1-2668, and the National Science Foundation under Grants IIS-2316306 and CNS-2330215.

References

[Abdulaal *et al.*, 2024] Ahmed Abdulaal, adamos hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2024.

[Afonja *et al.*, 2024] Tejumade Afonja, Ivaxi Sheth, Ruta Binkyte, Waqar Hanif, Thomas Ulas, Matthias Becker, and Mario Fritz. Llm4grn: Discovering causal gene regulatory networks with llms – evaluation through synthetic data generation, 2024.

[Ban *et al.*, 2023a] Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. Causal structure learning supervised by large language model, 2023.

[Ban *et al.*, 2023b] Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data, 2023.

[Binder *et al.*, 1997] John Binder, Daphne Koller, Stuart Russell, and Keiji Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244, 1997.

[Chen *et al.*, 2023] Lyuzhou Chen, Taiyu Ban, Xiangyu Wang, Derui Lyu, and Huanhuan Chen. Mitigating prior errors in causal structure learning: Towards llm driven prior knowledge, 2023.

[Chu *et al.*, 2023] Zhixuan Chu, Mengxuan Hu, Qing Cui, Longfei Li, and Sheng Li. Task-driven causal feature distillation: Towards trustworthy risk prediction. *arXiv preprint arXiv:2312.16113*, 2023.

[Cohrs *et al.*, 2023] Kai-Hendrik Cohrs, Emiliano Diaz, Vasileios Sitokonstantinou, Gherardo Varando, and Gustau Camps-Valls. Large language models for constrained-based causal discovery. In *AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?"*, 2023.

[Cohrs *et al.*, 2025] Kai-Hendrik Cohrs, Emiliano Diaz, Vasileios Sitokonstantinou, Gherardo Varando, and Gustau Camps-Valls. Large language models for causal hypothesis generation in science. *Mach. Learn.: Sci. Technol.*, 6(1):013001, January 2025.

[Darvariu *et al.*, 2024] Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. Large language models are effective priors for causal graph discovery. *arXiv preprint arXiv:2405.13551*, 2024.

[Drton and Maathuis, 2017] Mathias Drton and Marloes H Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4(1):365–393, 2017.

[Feng *et al.*, 2024] Tao Feng, Lizhen Qu, Niket Tandon, Zhuang Li, Xiaoxi Kang, and Gholamreza Haffari. From pre-training corpora to large language models: What factors influence llm performance in causal discovery tasks?, 2024.

[Gkountouras *et al.*, 2024] John Gkountouras, Matthias Lindemann, Phillip Lippe, Efstratios Gavves, and Ivan Titov. Language agents meet causality – bridging llms and causal world models, 2024.

[Glymour *et al.*, 2019] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[Heindorf *et al.*, 2020] Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. Causenet: Towards a causality graph extracted from the web. In *CIKM*, pages 1807–1814, 2020.

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

[Ji *et al.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[Jiang *et al.*, 2024] Haitao Jiang, Lin Ge, Yuhe Gao, Jianian Wang, and Rui Song. LLM4causal: Democratized causal tools for everyone via large language model. In *First Conference on Language Modeling*, 2024.

[Jin *et al.*, 2023a] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing causal reasoning in language models. In *NeurIPS*, 2023.

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

[Jiralerspong *et al.*, 2024] Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models, 2024.

[Kampani *et al.*, 2024] Shiv Kampani, David Hidary, Constantijn van der Poel, Martin Ganahl, and Brenda Miao. Llm-initialized differentiable causal discovery, 2024.

[Kuang *et al.*, 2020] Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, and Zhichao Jiang. Causal inference. *Engineering*, 6(3):253–263, 2020.

[Kıcıman *et al.*, 2023] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality, 2023.

[Le *et al.*, 2024] Hao Duong Le, Xin Xia, and Zhang Chen. Multi-agent causal discovery using large language models, 2024.

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

[Long *et al.*, 2023] Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

[Long *et al.*, 2024] Stephanie Long, Tibor Schuster, and Alexandre Piché. Can large language models build causal graphs?, 2024.

[Patel *et al.*, 2024] Parth Patel, Yu-Chiao Chiu, Yufei Hunag, and Jianqiu Zhang. Metaphorprompt - an analogical reasoning approach for extracting causal links from biological text. In *BCB*, 2024.

[Pearl, 1988] Judea Pearl. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

[Pearl, 2009] Judea Pearl. *Causal inference in statistics: An overview*. Cambridge University Press, 2009.

[Peters *et al.*, 2017] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. The MIT Press, Cambridge, MA, 2017.

[Petroni *et al.*, 2019] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases?, 2019.

[Scutari and Denis, 2014] Marco Scutari and Jean-Baptiste Denis. *Bayesian Networks: With Examples in R*. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis, 2014.

[Shen *et al.*, 2024] ChengAo Shen, Zhengzhang Chen, Dongsheng Luo, Dongkuan Xu, Haifeng Chen, and Jingchao Ni. Exploring multi-modal integration with tool-augmented llm agents for precise causal discovery, 2024.

[Shimizu *et al.*, 2011] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model, 2011.

[Sokolov *et al.*, 2024] Alik Sokolov, Fabrizzio Sabelli, Behzad faraz, Wuding Li, and Luis Seco. Towards automating causal discovery in financial markets and beyond. *SSRN Electronic Journal*, 01 2024.

[Spirtes *et al.*, 2001] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

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

[Tsamardinos *et al.*, 2006] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 2006.

[Tu *et al.*, 2023] Ruibo Tu, Chao Ma, and Cheng Zhang. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis, 2023.

[Vashishtha *et al.*, 2023] Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal inference using llm-guided discovery, 2023.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

[Willig *et al.*, 2022] Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality?, 2022.

[Yu *et al.*, 2024] Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Qingzhen Liu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. Improving causal reasoning in large language models: A survey, 2024.

[Zečević *et al.*, 2023] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023.

[Zhang *et al.*, 2024] Yuzhe Zhang, Yipeng Zhang, Yidong Gan, Lina Yao, and Chen Wang. Causal graph discovery with retrieval-augmented generation based large language models, 2024.

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

[Zheng *et al.*, 2018] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning, 2018.

[Zhou *et al.*, 2024] Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. Causalbench: A comprehensive benchmark for causal learning capability of llms, 2024.