Exploring Multi-Modal Data with Tool-Augmented LLM Agents for Precise Causal Discovery

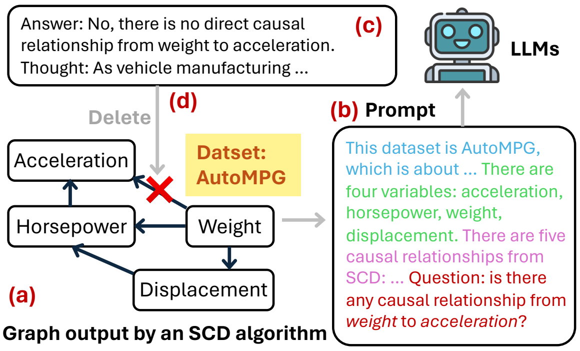
ChengAo Shen[1], **Zhengzhang Chen**[2*], **Dongsheng Luo**[3], **Dongkuan Xu**[4],
Haifeng Chen[2], **Jingchao Ni**[1*]
[1]University of Houston, [2]NEC Laboratories America,
[3]Florida International University, [4]North Carolina State University
[1]{cshen9, jni7}@uh.edu, [2]{zchen, haifeng}@nec-labs.com,
[3]dluo@fiu.edu, [4]dxu27@ncsu.edu

Abstract

Causal discovery is an imperative foundation for decision-making across domains, such as smart health, AI for drug discovery and AIOps. Traditional statistical causal discovery methods, while well-established, predominantly rely on observational data and often overlook the semantic cues inherent in cause-and-effect relationships. The advent of Large Language Models (LLMs) has ushered in an affordable way of leveraging the semantic cues for knowledge-driven causal discovery, but the development of LLMs for causal discovery lags behind other areas, particularly in the exploration of multi-modal data. To bridge the gap, we introduce MATMCD, a multi-agent system powered by tool-augmented LLMs. MATMCD has two key agents: a Data Augmentation agent that retrieves and processes modality-augmented data, and a Causal Constraint agent that integrates multi-modal data for knowledge-driven reasoning. The proposed design of the inner-workings ensures successful cooperation of the agents. Our empirical study across seven datasets suggests the significant potential of multi-modality enhanced causal discovery.[1]

1 Introduction

Identifying cause-and-effect relationships in complex systems is crucial for a variety of applications, including neuralgia diagnosis in medicine (Tu et al., 2019), protein pathway analysis in computational biology (Sachs et al., 2005), and root cause locating in microservice architectures (Wang et al., 2023a). These causal insights significantly benefit emerging fields such as smart health, AI-driven drug discovery, and AIOps. The process of discovering such relationships from observational data, known as *causal discovery* (Yu et al., 2025), typically generates a Directed Acyclic Graph (DAG). In this graph, edges represent the existence and direction of causal relationships between variables, as illustrated in Fig. 1(a). This DAG not only governs the data generation process but also enhances the understanding of inter-variable influences, serving as the foundation for many downstream decision-making tasks (Nguyen et al., 2023). As such, constructing accurate causal graphs is essential to the reliability of subsequent analyses.



Answer: No, there is no direct causal relationship from weight to acceleration. Thought: As vehicle manufacturing ...

This dataset is AutoMPG, which is about ... There are four variables: acceleration, horsepower, weight, displacement. There are five causal relationships from SCD: ... Question: is there any causal relationship from *weight* to *acceleration*?

Figure 1: An illustration of the general process of LLM based causal discovery: (a) a causal graph is estimated by an SCD algorithm; (b) a prompt is generated *w.r.t.* the graph; (c) LLMs reason about the causal structures; (d) the graph is refined based on the LLM response.

Conventional methods primarily rely on data-driven statistical causal discovery (SCD), which can be categorized as non-parametric (Spirtes and Glymour, 1991; Silander and Myllymäki, 2006; Huang et al., 2018; Xie et al., 2020) and semi-parametric (Shimizu et al., 2006, 2011; Zheng et al., 2018; Tu et al., 2022). These methods estimate causal relationships by analyzing the observational data of variables, but they overlook the semantic and contextual cues of the variables, resulting in suboptimal outcomes (Takayama et al., 2024).

Commonsense and domain knowledge are invaluable for identifying cause-and-effect relationships among semantically meaningful variables. In light of this, growing research attention has been drawn to elicit such information for causal discov-

[*] Corresponding authors.

[1] The code of MATMCD is available at https://github.com/D2I-Group/matmcd.

Findings of the Association for Computational Linguistics: ACL 2025, pages 636–660
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

ery. Large Language Models (LLMs), as praised by its astonishing reasoning ability drawing on extensive knowledge acquired from large-scale training (Brown et al., 2020; Achiam et al., 2023), now become a promising and cost-effective source of expert knowledge to aid causal discovery. For example, by merely prompting with variable names and some contextual cues, LLMs have been shown to infer meaningful causal relationships (Ban et al., 2023; Jiralerspong et al., 2024). More recently, hybrid approaches were introduced to combine LLMs with data-driven SCD algorithms, achieving enhanced accuracy in causal discovery (Takayama et al., 2024; Khatibi et al., 2024). However, most existing methods have yet to fully harness the potential of modern LLMs, particularly the agent systems built upon tool-augmented LLMs.

An LLM agent is typically equipped with memory, reasoning, planning, and access to external tools such as calculators, search engines, and code compilers, rendering it superior in problem-solving compared to vanilla LLMs (Yao et al., 2022). As a single-agent system may be hallucination-prone even with self-reflection (Li et al., 2023; Shinn et al., 2024), multi-agent systems were introduced, which rivals more advanced models by combining multiple weaker agents. Despite the fast progress, LLM agents for causal discovery remains underexplored. The most relevant effort to date introduced a multi-agent system with debating LLMs for synergistic causal discovery (Le et al., 2024). However, this approach has never explored the potential of *multi-modality* in data – a feature that agent systems are well-equipped to handle.

As Fig. 1 shows, hybrid methods for causal discovery typically prompt LLMs with the prior causal graph produced by some SCD algorithm, appended by some meta-data (*e.g.*, variable names, dataset titles) as contexts. However, these inputs may fall short in fully activating the reasoning ability of LLMs. Inspired by the observation that abundant semantic data from external sources, such as webs and logs, can serve as an *additional modality to the observational data (e.g., time series)* for improving prompts, we propose a Multi-Agent system with Tool-augmented LLMs for exploring Multi-modal data enhanced Causal Discovery (MATMCD).

Specifically, MATMCD is designed as a framework for refining causal graphs generated by SCD algorithms, involving two key agents: (1) a Data Augmentation Agent (DA-AGENT, §3.2), and (2) a Causal Constraint Agent (CC-AGENT, §3.3). Given a causal graph output by some SCD algorithm, DA-AGENT integrates meta-data (*i.e.*, variable names and dataset titles) and calls tools such as web search APIs or offline log lookup APIs for iterative, reflection-based retrieval of augmented (textual) data, which is summarized as a compact cue in a different modality from the graphs. Upon receiving the augmented data, CC-AGENT combines it with the topological structure of the prior causal graph to infer the causal relationships among variables. Both agents comprise multiple cooperative LLMs with well-crafted mechanisms to handle sub-tasks such as tool-calling, memorizing, reasoning, and summarization. Retrieval Augmented Generation (RAG) (Lewis et al., 2020) components are used where efficient memory is essential. In our experiments, we compared MATMCD with state-of-the-art (SOTA) baseline methods across five benchmark datasets and two public AIOps datasets of microservice systems. The results demonstrate substantial improvements in causal discovery by incorporating multi-modal data. The main contributions of this work are summarized as follows:

- We propose to explore the problem of multi-modal data enhanced causal discovery via LLM agents, which is significant yet less studied.
- We introduce MATMCD, a novel framework of multi-agent, as a testbed for assessing the effectiveness of multi-modal data in causal discovery.
- We perform extensive experiments on a variety of datasets, where MATMCD reduces causal discovery errors (NHD) by up to 66.7% and improves root cause locating (MAP@10) by up to 83.3% over the best baselines, suggesting the potential of multi-modal data in causal discovery.

2 Related Work

Causal Discovery Methods are mostly conventional data-driven SCD algorithms, including non-parametric methods (Spirtes and Glymour, 1991; Chickering, 2002; Silander and Myllymäki, 2006; Huang et al., 2018; Xie et al., 2020) and semi-parametric methods (Shimizu et al., 2006; Hoyer et al., 2008; Shimizu et al., 2011; Zheng et al., 2018; Rolland et al., 2022; Tu et al., 2022). These methods rely on observational data as input but cannot leverage the semantics of variables. Recently, knowledge-driven methods have been found promising for causal discovery. Some early efforts use LLMs by simply prompting variable names and dataset titles (Kıcıman et al., 2023; Zečević et al., 2023; Chen et al., 2024a; Jiralerspong et al., 2024). Then hybrid approaches that integrate SCD algorithms with LLMs were introduced (Ban et al., 2023; Vashishtha et al., 2023; Khatibi et al., 2024; Takayama et al., 2024) and found to be more effective than pure LLM-based methods. More recently, a multi-agent system-based approach was proposed (Le et al., 2024) to explore the impacts of debating LLMs. However, none of these methods has explored the potential of multi-modal data in the LLM-based causal discovery process.

LLM Agents are typically equipped with plan, memory and tools. Planning can use techniques such as CoT (Wei et al., 2022), ReAct (Yao et al., 2022), and Reflexion (Shinn et al., 2024). Tools endow agents with the ability to interact with environments. MRKL (Karpas et al., 2022), Toolformer (Schick et al., 2024), Function Calling (OpenAI, 2024), and HuggingGPT (Shen et al., 2024) exemplify paradigms that integrate tools for problem-solving. For complex tasks, multi-agent systems are promising. The primary categories of multi-agent systems include cooperative agents (Qian et al., 2023; Chen et al., 2024b), competitive agents (Zhao et al., 2023) and debating agents (Li et al., 2023; Liang et al., 2023; Xiong et al., 2023). In this work, we investigate a cooperative multi-agent system where the agents are coordinated to enhance a solution towards a shared goal.

3 Methodology

Fig. 2 is an overview of the proposed MatMcd system, which has four key components: (1) Causal Graph Estimator (§3.1); (2) Data Augmentation Agent (DA-Agent, §3.2); (3) Causal Constraint Agent (CC-Agent, §3.3); and (4) Causal Graph Refiner (§3.4). Next, we will first introduce some notations and then elaborate on each of the components in an order subject to the flow of data.

Notations. Suppose there is a set of n variables $\mathcal{V} = \{v_1, ..., v_n\}$ (*e.g.*, 4 variables in Fig. 1(a)), each variable v_i is associated with a set of observed data samples $\mathbf{v}_i = \{v_{i1}, ..., v_{im}\}$ where m is the number of samples. The observational data could be random samples or a length-m time series emitted by each variable, depending on the application. In many cases, meta-data is available. In this work, we assume a minimal set of meta-data $\mathcal{D} = \{\mathsf{s}, \mathcal{Z}\}$ where s is a descriptive title of the dataset (*e.g.*, "AutoMPG" in Fig. 1(a)) and $\mathcal{Z} = \{\mathsf{z}_1, ..., \mathsf{z}_n\}$ includes the descriptive name z_i of each variable v_i in $\mathcal{V}$ (*e.g.*, "Acceleration" in Fig. 1(a)).

The Task. Based on $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ and $\mathcal{D}$, we want to construct a DAG, $\mathbf{G} = (\mathcal{V}, \mathcal{E})$, where each variable v_i in $\mathcal{V}$ is a node in the graph, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of *directed* edges with $(v_i, v_j) \in \mathcal{E}$ signifying a causal relationship from v_i to v_j. The goal of this work is to infer accurate causal relationships in $\mathcal{E}$ among the set of n variables in $\mathcal{V}$.

3.1 Causal Graph Estimator

Similar to the hybrid approaches (Takayama et al., 2024; Khatibi et al., 2024), Causal Graph Estimator serves as an initializer of causal graph and is built upon data-driven SCD algorithms, with the aim of estimating an initial causal graph $\mathbf{G}_0 = (\mathcal{V}, \mathcal{E}_0)$ purely from the observational data $\{\mathbf{v}_1, ..., \mathbf{v}_n\}$ without accessing any other information. Here $\mathcal{V}$ is not subscript as it will be kept intact throughout the proposed framework and our focus is on the alteration of the causal relationships in $\mathcal{E}$ for accurate causal discovery.

Our framework is flexible to the choice of SCD algorithms. In this work, we investigate the feasibility of employing three widely used algorithms, each of which is a representative of a category: (1) Constraint-based method – Peter-Clark (PC) algorithm (Spirtes and Glymour, 1991) which is non-parametric; (2) Score-based method – Exact Search (ES) algorithm (Yuan and Malone, 2013) which is non-parametric; and (3) Constrained functional causal models – DirectLiNGAM (Shimizu et al., 2011) which is semi-parametric. We leave the exploration of other SCD algorithms in our future work as it is out of the scope of this work. Inspired by the recent LLM prompting techniques for graphs (Wang et al., 2024), the output edges $\mathcal{E}_0$ of the SCD algorithm will be embedded as an adjacency list in the prompt generated by the prompt builder module of the CC-Agent (§3.3) in Fig. 2, which will also integrate the semantics-rich data modality retrieved by the DA-Agent (§3.2).

3.2 Data Augmentation Agent (DA-Agent)

The goal of DA-Agent is to retrieve semantics-rich contextual data pertinent to the initial causal graph $\mathbf{G}_0$, such as web documents and log files about the variables, as an additional modality to observational data for prompting CC-Agent (§3.3). As illustrated in Fig. 2(b), DA-Agent comprises a Search LLM and a Summary LLM.

Search LLM. The Search LLM has access to a

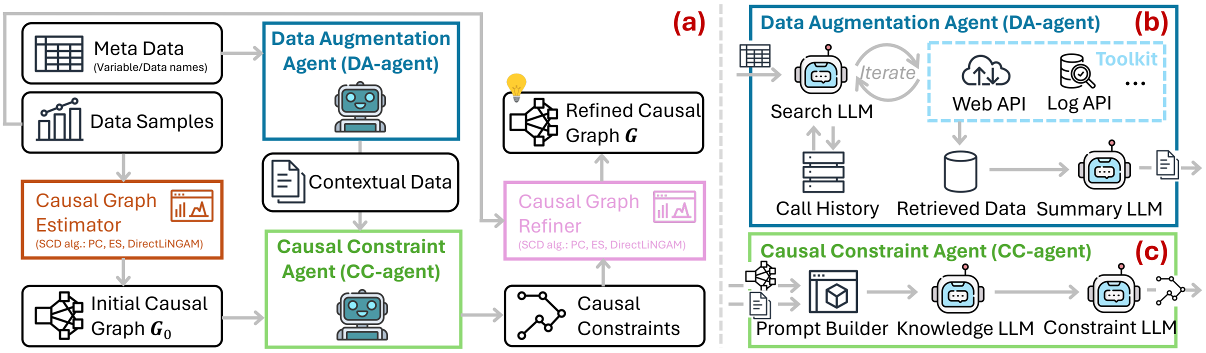


Figure 2: An illustration of the MATMCD framework: (a) an overview of the framework, (b) the inner working of DA-AGENT, and (c) the inner working of CC-AGENT.

set of tools for data search. In this work, we focus on a web search API as a general tool for retrieving contextual data from external sources, and a log lookup API for applications where a domain-specific database is available such as the process logs in root cause analysis for microservice systems in AIOps (Zheng et al., 2024a). The DA-AGENT is flexible to the toolkit and is extensible for a wide scenarios by including other application-specific tools such as Wikipedia API and code lookup API, which we leave for future exploration.

As shown in Fig. 2, upon receiving the meta-data $\mathcal{D} = \{\mathsf{s}, \mathcal{Z}\}$ about the causal graph, the Search LLM first checks its *calling history memory* to decide whether to initiate a new tool call. If a new call is needed, the Search LLM invokes a search tool API to retrieve additional data using a prompt that includes the dataset title s and variable names $\mathsf{z}_1, ..., \mathsf{z}_n$. This search action is then recorded in the memory for future reference. Since our focus is on search tools and the search action involves generating a query, the generated query is added to the memory. In subsequent iterations, all previously recorded queries are examined to prevent redundant queries. This process continues iteratively until the Search LLM determines that no further tool calls are necessary, terminating the loop.

To enable this iterative search, the prompt is designed based on self-reflection techniques (Shinn et al., 2024; Madaan et al., 2024), where the LLM assesses whether additional queries are needed based on the comprehensiveness of the historical queries. The loop terminates when the LLM concludes with a "No query needed" response. Compared to single-round searches, this iterative process proves crucial for retrieving relevant and comprehensive data, especially in domains where variable-specific information is challenging to locate (*e.g.*, medicine). However, for lookup APIs, iteration is unnecessary. Thus, the retrieved data through these tools constitutes an additional textual modality for the causal graph. The prompt used for the Search LLM is provided in Appendix C.1.

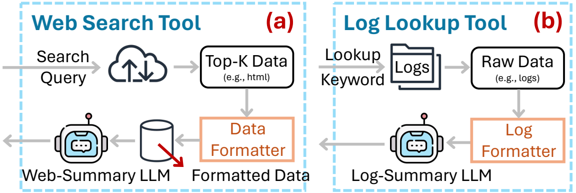


Figure 3: The search tool preparation in DA-AGENT for (a) web search tool, and (b) log lookup tool.

Tool Preparation. Fig. 3 illustrates our preparation of the Web Search tool and the Log Lookup tool. In the former, we employ Google search API where the query is generated by the Search LLM as aforementioned. The retrieved top webpages will be de-formatted (*e.g.*, removing HTML tags) by a data formatter and the resultant plain docs will be stored in a memory. Then a *Web-Summary LLM* is employed to summarize the docs into a concise description. In contrast, the Log Lookup tool uses exact lookup, *i.e.*, with a variable name as the keyword, its corresponding log can be retrieved directly. Thus the memory can be removed and the retrieved log, which still needs de-formatting (*e.g.*, removing log templates) and could be lengthy, will be summarized by a *Log-Summary LLM*.

Summary LLM. The data retrieved by the Search LLM is iteratively added to the *Retrieved Data Memory*, as shown in Fig. 2(b). Upon loop termination, a *Summary LLM* summarizes the retrieved data into three types of cues: (1) description of the dataset; (2) description of each variable in the graph; and (3) relationships between the variables. Since the size of the retrieved data from iterative searches may exceed the LLM's context window, we adopt an efficient summarization approach using RAG (Lewis et al., 2020). The retrieved data is divided into indexed document chunks, implemented with `LlamaIndex` (Liu, 2022) using `text-embedding-ada-002` for chunk indexing and Maximum Inner Product Search (MIPS) for retrieving relevant chunks. An example summary is provided in Appendix D.1. The resulting summary serves as *contextual data* for the initial causal graph $\mathbf{G}_0$ to prompt CC-AGENT.

Remark. Since there is a risk of retrieving web content that may leak ground truth causal graphs for certain datasets, we conducted thorough screening in our RAG implementation of the *Retrieved Data Memory* to prevent such information leaks.

3.3 Causal Constraint Agent (CC-AGENT)

In addition to the external knowledge, our method leverages the factual knowledge stored in LLMs, acquired during pre-training. To achieve this, CC-AGENT is designed based on the Two-Stage Prompting framework of zero-shot Chain-of-Thought (ZS-COT) (Kojima et al., 2022). First, a prompt builder integrates $\mathbf{G}_0$, represented as an adjacency list, with contextual data from DA-AGENT to prompt a *Knowledge LLM*. The Knowledge LLM is tasked with explaining each (non-)existing causal relationship in the initial causal graph $\mathbf{G}_0$ based on the contextual data and its own knowledge. These explanations, which could either support or refute the causal relationships, are used to prompt a *Constraint LLM* in the second stage to draw a conclusion on the existence of each relationship (*i.e.*, "Yes"/"No"). To address potential uncertainty in the conclusions, we adopt the Top-K-Guess technique (Tian et al., 2023) to elicit verbal confidence. This approach, found to be more reliable than sampling-based likelihood estimation (Xiong et al., 2024), quantitatively evaluates the likelihood of each causal relationship. Among the Top-K guesses, the most confident one is selected as the final conclusion for each causal relationship.

3.4 Causal Graph Refiner

To ensure the final causal graph $\mathbf{G}$ is acyclic, the edge set $\mathcal{E}_0$ is not directly modified based on the (non-)existence constraints from CC-AGENT. Instead, the SCD algorithm used in the Causal Graph Estimator (§3.1) is rerun with these constraints imposed to generate a new causal graph $\mathbf{G}$. Specifically, upon receiving the (non-)existence constraints, a *constraint matrix* $\mathbf{C} \in \mathbb{R}^{n\times n}$ is constructed, where $\mathbf{C}_{ij} = 1$ if CC-AGENT indicates a causal effect from v_i to v_j, and $\mathbf{C}_{ij} = 0$ otherwise.

The most representative SCD algorithms, including PC, ES, and DirectLiNGAM, are designed to incorporate such a constraint matrix $\mathbf{C}$ as input alongside observational data $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$. This ensures that the generated causal graph complies with the constraints in $\mathbf{C}$ to varying extents, enabling the production of a refined causal graph $\mathbf{G}$ that is both consistent with the imposed constraints and a directed acyclic graph (DAG).

Remark. Compared to the existing approaches that utilize LLMs for causal discovery (Takayama et al., 2024; Khatibi et al., 2024), the key novelty of the proposed MATMCD lies in its exploration of DA-AGENT for multi-modal data enhanced causal discovery. An algorithmic summary of the proposed workflow is provided in Appendix A.

4 Experiments

In this section, we first compare MATMCD with SOTA methods on benchmark datasets. Then we evaluate MATMCD for a root cause analysis task on real-life enterprise microservice system datasets.

4.1 Experimental Setup

Benchmark Datasets. To be comprehensive, we use 5 benchmark datasets covering both continuous variables and discrete variables. For the former, following (Takayama et al., 2024; Le et al., 2024), we adopt (1) `AutoMPG` (Quinlan, 1993), which has five variables concerning city-cycle fuel consumption in miles per gallon, each variable has a length-392 time series; (2) `DWDClimate` (Mooij et al., 2016), which has six variables pertinent to observations from weather stations in Deutscher Wetterdienst, each variable has a length-350 time series; and (3) `SachsProtein` (Sachs et al., 2005), which has eleven variables measuring the expression level of different proteins and phospholipids in human cells, each variable has a length-7,466 time series. For the latter, following (Long et al., 2023; Jiralerspong et al., 2024), we adopt (4) `Asia` (Lauritzen and Spiegelhalter, 1988), which has eight variables relevant to lung disease diagnosis, each variable has 1,000 discrete samples; and (5) `Child` (Spiegelhalter, 1992), which has twenty variables regarding congenital heart disease in newborn babies, each variable has 1,000 discrete samples.

In `AutoMPG`, `DWDClimate`, and `SachsProtein`,

Method	AutoMPG					DWDClimate					SachsProtein				
	Prc ↑	F1 ↑	FPR ↓	SHD ↓	NHD ↓	Prc ↑	F1 ↑	FPR ↓	SHD ↓	NHD ↓	Prc ↑	F1 ↑	FPR ↓	SHD ↓	NHD ↓
PC	0.11	0.14	0.40	8	0.32	0.14	0.15	0.20	9	0.25	0.38	**0.44**	0.15	24	0.19
Exact Search	0.25	0.30	0.30	6	0.24	0.45	0.58	0.20	6	0.16	0.18	0.23	0.26	31	0.25
DirectLiNGAM	0.11	0.14	0.40	8	0.32	0.16	0.22	0.33	10	0.27	0.27	0.36	0.25	29	0.23
MAC*	-	-	-	4	0.16	-	-	-	6	0.19	-	-	-	21	0.19
Efficient-CDLMs	0.66	0.50	0.05	4	0.16	0.33	0.33	0.13	8	0.22	0.33	0.09	**0.02**	20	0.16
SCD-LLM	0.57	0.66	0.15	3	0.12	0.33	0.22	0.06	7	0.19	0.04	0.05	0.19	29	0.23
ReAct	0.50	0.54	0.15	4	0.16	0.60	0.40	0.06	6	0.16	0.04	0.04	0.20	29	0.23
LLM-KBCI	0.57	0.66	0.15	3	0.12	0.50	0.40	0.06	6	0.16	0.13	0.14	0.19	27	0.22
LLM-KBCI-RA	0.57	0.55	0.15	3	0.12	**0.75**	0.60	**0.03**	**4**	**0.11**	0.08	0.09	0.20	30	0.24
LLM-KBCI-RE	0.50	0.61	0.20	4	0.16	0.50	0.40	0.06	6	0.16	0.09	0.10	0.18	28	0.23
MatMcd	**1.00**	**0.88**	**0.00**	**1**	**0.04**	**0.75**	**0.88**	**0.03**	**4**	**0.11**	**0.50**	0.42	0.06	**17**	**0.14**
MatMcd-RE	0.57	0.66	0.15	3	0.12	0.50	0.40	0.06	6	0.16	0.31	0.31	0.12	21	0.17

Table 1: Comparison of different causal discovery methods on datasets with continuous variables. ↑ indicates larger score is better. ↓ indicates smaller score is better. * indicates numbers are adopted from the papers of the methods.

Method	Asia					Child				
	Prc ↑	F1 ↑	FPR ↓	SHD ↓	NHD ↓	Prc ↑	F1 ↑	FPR ↓	SHD ↓	NHD ↓
PC	0.50	0.50	0.07	6	0.09	0.30	0.34	0.06	24	0.06
Exact Search	0.50	0.42	0.05	6	0.09	0.35	0.28	**0.02**	**19**	**0.04**
DirectLiNGAM	0.28	0.36	0.17	11	0.17	0.29	0.36	0.07	33	0.08
Efficient-CDLMs	0.57	0.53	0.05	7	0.10	0.21	0.20	0.048	37	0.09
SCD-LLM	0.60	0.40	**0.03**	5	0.07	**0.56**	**0.54**	**0.02**	**19**	**0.04**
ReAct	0.40	0.30	0.05	6	0.09	**0.56**	**0.54**	**0.02**	**19**	**0.04**
LLM-KBCI	0.42	0.40	0.07	6	0.09	0.48	0.50	0.03	20	0.05
LLM-KBCI-RA	0.33	0.28	0.07	7	0.10	0.44	0.46	0.04	21	0.05
LLM-KBCI-RE	0.28	0.26	0.08	7	0.10	0.40	0.42	0.04	22	0.05
MatMcd	0.50	0.42	0.05	6	0.09	0.48	0.50	0.03	20	0.05
MatMcd-RE	**0.66**	**0.57**	**0.03**	**4**	**0.06**	**0.56**	**0.54**	**0.02**	**19**	**0.04**

Table 2: Comparison of different causal discovery methods on datasets with discrete variables.

ground truth causal graphs constructed by experts are available for evaluation purpose. For Asia and Child, since the observational data of the variables are sampled from a Bayesian network, the prior conditional probabilities among the variables establish the ground truth causal graphs.

Baselines. We compare MatMcd with the most relevant SOTA methods on causal discovery, including (1) statistical causal discovery: Peter-Clark (PC) algorithm (Spirtes and Glymour, 1991), Exact Search (ES) algorithm (Yuan and Malone, 2013), and DirectLiNGAM (Shimizu et al., 2011); (2) LLM-based causal discovery that only uses LLMs to infer causal relationships: Efficient-CDLMs (Jiralerspong et al., 2024), which employs a BFS-based LLM prompting for efficient causal graph construction, and MAC (Le et al., 2024), which uses Debating LLMs for building a multi-agent system; and (3) Hybrid approaches that refine an SCD causal graph by LLMs: SCD-LLM, which uses a single LLM upon the SCD output, ReAct (Yao et al., 2022), which interleaves reasoning and search tool usage when refining SCD graphs, LLM-KBCI (Takayama et al., 2024), which uses ZSCOT two-stage prompting for refining causal graphs.

Moreover, we apply ReAct framework for LLM-KBCI to enable alternate reasoning and tool usage and name this baseline as LLM-KBCI-RA. We also introduce Top-K Guess reasoning (Tian et al., 2023) (with K=2) for verbal calibration on LLM-KBCI and name this variant as LLM-KBCI-RE.

For our method, we consider two major variants. The first asks for a single answer from the CC-Agent (§3.3), *i.e.*, K=1 in the Top-K Guess reasoning, named as MatMcd. The second uses K=2 in the Top-K Guess reasoning, which we name as MatMcd-RE. Additionally, we perform extensive ablation analysis on other variants of MatMcd in Table 3 to evaluate its design choices.

Implementation. By default, we use GPT-4o mini with temperature 0.5 as the base LLM for all LLM-based and Hybrid methods, as it was found very

Method	AutoMPG					DWDClimate					SachsProtein				
	Prc ↑	F1 ↑	FPR ↓	SHD ↓	NHD ↓	Prc ↑	F1 ↑	FPR ↓	SHD ↓	NHD ↓	Prc ↑	F1 ↑	FPR ↓	SHD ↓	NHD ↓
MatMcd	**1.00**	**0.88**	**0.00**	**1**	**0.04**	**0.75**	**0.88**	**0.03**	**4**	**0.11**	0.50	0.42	0.06	17	0.14
(a) Iter.→Single search	0.66	0.72	0.10	2	0.08	0.66	0.44	**0.03**	5	0.13	0.38	0.22	0.04	17	0.14
(b) w/o Knowledge LLM	0.50	0.61	0.20	4	0.16	0.66	0.66	0.06	**4**	**0.11**	0.19	0.20	0.16	26	0.21
(c) PC→ES	0.33	0.42	0.30	6	0.24	0.50	0.50	0.10	5	0.13	0.26	0.28	0.16	25	0.20
(c) PC→DirectLiNGAM	0.16	0.18	0.25	6	0.24	0.16	0.22	0.23	9	0.25	0.19	0.22	0.20	27	0.22
(d) LLM→GPT4	0.57	0.66	0.15	3	0.12	0.60	0.54	0.06	5	0.13	**0.58**	**0.45**	**0.04**	**15**	**0.12**
(d) LLM→Llama3.1-8B	0.37	0.46	0.25	5	0.20	0.33	0.22	0.06	7	0.19	0.16	0.18	0.20	29	0.23
(d) LLM→Llama3.1-70B	0.50	0.54	0.15	4	0.16	0.40	0.36	0.10	7	0.19	0.26	0.26	0.13	24	0.19
(d) LLM→Gemma2-9B	0.37	0.46	0.25	5	0.20	0.16	0.16	0.16	8	0.22	0.16	0.18	0.20	29	0.23
(d) LLM→Ministral-7B	0.60	0.60	0.10	4	0.16	0.33	0.33	0.13	7	0.19	0.21	0.18	0.10	25	0.20

Table 3: Ablation analysis of the proposed MatMcd method on benchmark datasets.

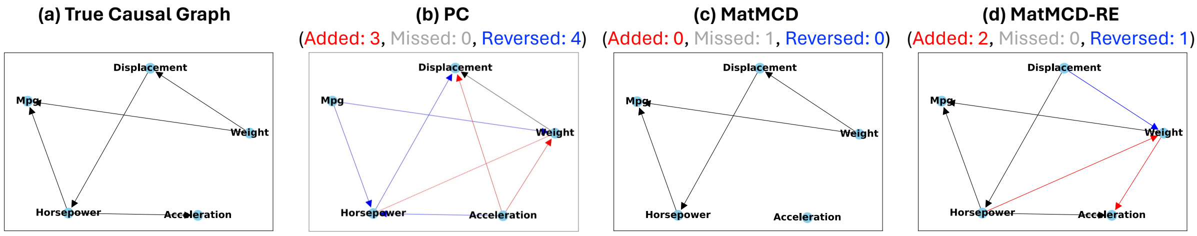
performant by the existing works (Le et al., 2024). For all Hybrid approaches, PC is used as the base SCD algorithm. Also, we evaluate our MatMcd by switching the LLM with GPT-4, Llama-3.1-8B, Llama-3.1-70B, Mistral-7B and Gemma2-9B, and switching the SCD algorithm with ES and DirectLiNGAM in our ablation analysis. All SCD algorithms were implemented with causal-learn (Zheng et al., 2024b). ReAct and RAG frameworks were implemented with LlamaIndex (Liu, 2022). For all baselines, we used their official code when available. Since MAC's code is unavailable, we report its results from the original paper, which are only available on AutoMPG, DWDClimate, and SachsProtein datasets.

Evaluation Metrics. Following (Kıcıman et al., 2023; Takayama et al., 2024; Khatibi et al., 2024; Le et al., 2024), we employ the widely used metrics including precision (Prc), F1-score (F1), FPR, structural Hamming distance (SHD), and normalized Hamming distance (NHD) for gauging the difference between the predicted causal graphs and the ground truth graphs. Prc and F1 measure the accuracy, thus a larger value is better. FPR, SHD, and NHD measure the errors/differences, hence a smaller value is better.

4.2 Experimental Results

Causal Discovery. Table 1 and 2 summarize the results on the benchmark datasets. From the tables, we have several observations: (1) Methods involving LLMs generally outperform SCD algorithms in most cases except for SachsProtein and Asia datasets which pertain to biomedicine. It demonstrates the great potential of LLMs' commonsense knowledge acquired by pre-training for the task of causal discovery, but meanwhile draws to our attention on their short of domain-specific knowledge. (2) Hybrid approaches outperform LLM-based baselines in most cases, indicating using SCD outputs as a prior for LLMs to reference has the benefits of complementing their causal effects related knowledge. (3) Baseline methods that can leverage tools for retrieving external data, *i.e.*, ReAct and LLM-KBCI-RA, sometimes outperform their counterparts, suggesting the potential of the augmented knowledge, but calling for a better way of retrieving and using external data. (4) The proposed MatMcd(-RE) achieved the best overall performance considering no baseline methods consistently performs well across all datasets. In particular, MatMcd(-RE) helps alleviate the hallucination problem of other LLM baselines on the biomedical SachsProtein and Asia datasets via data augmentation. The results highlight the challenge of the existing LLM baselines in inferring causal effects solely by meta-data (*i.e.*, node names, data titles) and validate the effectiveness of the proposed design of a multi-agent system for exploring multi-modality enhanced causal discovery. Finally, (5) MatMcd-RE outperform MatMcd on the Bayesian datasets Asia and Child, which could be a result of the better calibrated likelihoods by Top-K Guess reasoning for the probabilistic datasets.

Ablation Study. Table 3 presents our ablation analysis using three datasets. In the table, MatMcd is our original model. In (a), we aim to study the usefulness of the iterative search in DA-Agent. To this end, we replaced the iterative search with a single-round search template (Appendix C.3). In (b), we assess the effectiveness of the Knowledge LLM in CC-Agent. This is achieved by removing it from CC-Agent, which is equivalent to merge Knowledge LLM and Constraint LLM into a single LLM (the template is in Appendix C.4). In (c), we evaluate how MatMcd could generalize to differ-

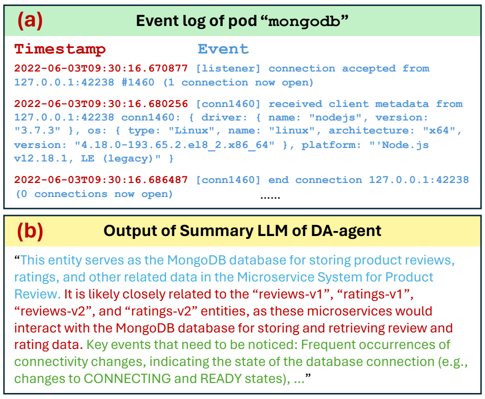

(a) True Causal Graph

(b) PC
(Added: 3, Missed: 0, Reversed: 4)

(c) MatMCD
(Added: 0, Missed: 1, Reversed: 0)

(d) MatMCD-RE
(Added: 2, Missed: 0, Reversed: 1)

Figure 4: Causal graph visualization for AutoMPG dataset. Red (Blue) arrows indicate edges that are wrongly added (reversed). Each sub-caption includes the number of wrongly added, missed, and reversed edges for comparison.


(a) Event log of pod "mongodb"

```
Timestamp                  Event
2022-06-03T09:30:16.670877 [listener] connection accepted from
127.0.0.1:42238 #1460 (1 connection now open)

2022-06-03T09:30:16.680256 [conn1460] received client metadata from
127.0.0.1:42238 conn1460: { driver: { name: "nodejs", version:
"3.7.3" }, os: { type: "Linux", name: "linux", architecture: "x64",
version: "4.18.0-193.65.2.el8_2.x86_64" }, platform: "'Node.js
v12.18.1, LE (legacy)" }

2022-06-03T09:30:16.686487 [conn1460] end connection 127.0.0.1:42238
(0 connections now open)                  ......
```

(b) Output of Summary LLM of DA-agent

"This entity serves as the MongoDB database for storing product reviews, ratings, and other related data in the Microservice System for Product Review. It is likely closely related to the "reviews-v1", "ratings-v1", "reviews-v2", and "ratings-v2" entities, as these microservices would interact with the MongoDB database for storing and retrieving review and rating data. Key events that need to be noticed: Frequent occurrences of connectivity changes, indicating the state of the database connection (e.g., changes to CONNECTING and READY states), ..."

Figure 5: An illustration of (a) log of events, and (b) its summary provided by DA-AGENT.

Method	MAP@5 ↑	MAP@10 ↑	MRR ↑	RK (P) ↓	RK (C) ↓
PC	0.0%	25.0%	0.14	5	13
Efficient-CDLMs	0.0%	0.0%	0.10	10	10
SCD-LLM	0.0%	25.0%	0.14	5	13
ReAct	0.0%	25.0%	0.14	5	12
LLM-KBCI	10.0%	30.0%	0.16	4	13
LLM-KBCI-RA	0.0%	25.0%	0.14	5	12
LLM-KBCI-RE	10.0 %	30.0%	0.16	4	13
MATMCD	**30.0%**	**55.0%**	**0.32**	**2**	7
MATMCD-RE	20.0%	**55.0%**	0.25	3	**6**

Table 4: Comparison of different causal discovery methods for the RCA task. RK (P) and RK (C) are the ranks of the root causes in the Product Review dataset and Cloud Computing dataset respectively, predicted by different methods.

ent SCD algorithms by switching PC with ES and DirectLiNGAM. In (d), we assess the impact of different LLMs on causal discovery.

From Table 3(a), we observe iterative search is better than single-round search, as the latter may use biased query and have difficulty in producing comprehensive augmented data. In Appendix D.1, we summarized some queries iteratively generated by DA-AGENT. From Table 3(b), we can see including a separate Knowledge LLM for explaining causal relationships is better than merging it with Constraint LLM, suggesting their distinct roles in CC-AGENT. In Table 3(c), using other SCD algorithms than PC generally degrades the performance, possibly because the constraint-based design of PC leads to better use of the constraints produced by LLMs. We also observe that, with MATMCD, the performance of ES and DirectLiNGAM were slightly enhanced compared to the counterparts without MATMCD in Table 1. Intriguingly, Table 3(d) shows the default GPT-4o mini is the most robust across different datasets, and GPT4 is better than the open source LLMs. This is similar to the findings in (Le et al., 2024), where MAC with GPT-4o mini performed the best. Compared to PC in Table 1, all LLM variants in Table 3(d) improve performance, suggesting the effectiveness in leveraging the LLMs. Moreover, on the biological dataset SachsProtein, a larger model GPT4 further boost the causal discovery performance, likely due to its better alignment with the specific domain.

Graph Visualization. In Fig. 4, we investigate how MATMCD(-RE) refines the causal graphs generated by SCD algorithms by visualizing the true causal graph of the AutoMPG dataset, along with the causal graphs produced by PC, MATMCD and MATMCD-RE. The visualization results on other datasets are deferred to Appendix §B.3. Compared to PC, MATMCD(-RE) yields graphs that resemble the true causality better, with less wrongly added, missed, and reversed edges, suggesting the advantage of leveraging LLM agents' ability to impose multi-modal data empowered knowledge-driven constraints for screening erroneous edges. In contrast, PC alone relies solely on observational data thus is more error-prone. Moreover, LLM-provided knowledge offers precise guidance in establishing causal directions, resulting in fewer reversed edges on most of the datasets.

4.3 Case Study: Root Cause Analysis (RCA)

Next, we evaluate MATMCD on AIOps datasets collected from Product Review (PR) and Cloud Computing (CC) microservice systems (Zheng et al., 2024a). The PR dataset has 216 variables (*i.e.*, system pods), each associated with a multivariate time series containing 6 metrics (*e.g.*, CPU, memory usage) of length 131,329. The CC dataset has 168 variables, each with a multivariate time series of 7 metrics and a length of 109,351. In both datasets, each variable also has a log recording its historical events. Fig. 5(a) illustrates a log snippet. Through the log lookup tool (§3.2), these logs can be leveraged as an additional data modality by MATMCD for enhanced causal discovery. Fig. 5(b) demonstrates that the Summary LLM of DA-AGENT can effectively interpret and summarize the log data.

Since these datasets provide root causes of system failures identified by domain experts, but lack ground truth causal graphs, we use them to evaluate the root cause analysis (RCA) performance based on the causal graphs produced by different methods. Following (Wang et al., 2023b), if running a random walk with restart (RWR) on a causal graph can top-rank the root cause among all variables, it reflects the quality of the causal graph. Therefore, we use widely adopted metrics, including Mean Average Precision@K (MAP@K), with K set to 5 and 10, and Mean Reciprocal Rank (MRR), to assess the overall ranking performance on both datasets (Zheng et al., 2024a). More details on MAP@K and MRR can be found in Appendix B.2.

Following (Wang et al., 2023b), we preprocessed the datasets by filtering out irrelevant pods using an Extreme Value Theory-based approach (Siffer et al., 2017) before applying causal discovery methods. Table 4 summarizes the results, where we also report the specific ranking ("RK") of the ground truth root causes for each dataset. In Table 4, ES and DirectLiNGAM are excluded as the focus is how different LLM-based methods can improve their base SCD algorithm, *i.e.*, PC, for the downstream task. From Table 4, by leveraging the log modality, MATMCD(-RE) significantly improves the accuracy of root cause locating on both datasets, with an 83.3% relative improvement over the best baseline on MAP@10. This highlights the potential of integrating prevalent log data in microservice systems for RCA and AIOps.

5 Conclusion

In this paper, we explored multi-modal data in causal discovery via devising MATMCD with tool-augmented LLMs, which has a well-crafted mechanism for integrating textual and observational data. The experiments not only validate the effectiveness of our innovative method but also set a groundwork for delving into multi-modal causal discovery.

6 Limitations

The research of LLM based causal discovery is an emergent area undergoing active explorations. This work pioneers the use of tool-augmented LLM agents, but the proposed method shares similar limitations of most of the existing approaches.

First, the examination of the causal relationships in a causal graph is pair-based, *i.e.*, each causal relationship (v_i, v_j) for $1 \leq i < j \leq n$ needs to be prompted to an LLM (in our case, CC-AGENT) for assessing its correctness. This may lead to scalability issues on large graphs. This is also reflected in the sizes of the benchmark datasets widely used in this area. A promising way to improve the scalability of such edge-based QA paradigm is to leverage the DAG structure of causal graphs and use breadth first search (BFS) style prompts as introduced in (Jiralerspong et al., 2024). The proposed MATMCD is completely compatible with this prompting technique, thus has the potential to achieve an $O(n)$ complexity. Second, similar to other hybrid approaches that combine SCD algorithm and LLM agents, the proposed MATMCD necessitates a base SCD algorithm that can effectively incorporates the constraints produced by LLMs. We observed this from the better results of using PC compared to use either ES or DirectLiNGAM in Table 3. This is worth further study with a comprehensive comparison of using more constraint-based SCD algorithms and non-constraint-based variants. Third, as a knowledge-driven method, meta-data with semantically meaningful variables is needed for retrieving useful knowledge about causal relationships. For domains where variables are semantically meaningless or meta-data are private, such methods may have limited impact.

In addition to the aforementioned common limitations, this work has several specific limitations. First, this work uses web data and logs as the augmented data modality of focus, more modalities such as codes, images and audio signals in certain domains remain further study. Second, tools beyond the web search APIs and log lookup APIs in the proposed DA-AGENT, such as Wikipedia APIs and code lookup APIs, are worth exploring. The toolkit in the DA-AGENT has the flexibility and extensibility to support further research.

References

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

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. 2023. Causal structure learning supervised by large language model. *arXiv preprint arXiv:2311.11689*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*.

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

Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. 2024b. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems? In *ICRA*.

David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(Nov):507–554.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. 2008. Nonlinear causal discovery with additive noise models. *NeurIPS*.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. 2018. Generalized score functions for causal discovery. In *KDD*.

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

Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, et al. 2022. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445*.

Elahe Khatibi, Mahyar Abbasian, Zhongqi Yang, Iman Azimi, and Amir M Rahmani. 2024. Alcm: Autonomous llm-augmented causal discovery framework. *arXiv preprint arXiv:2405.01744*.

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *NeurIPS*.

Steffen L Lauritzen and David J Spiegelhalter. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. Ser. B Methodol.*, 50(2):157–194.

Hao Duong Le, Xin Xia, and Zhang Chen. 2024. Multi-agent causal discovery using large language models. *arXiv preprint arXiv:2407.15073*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *NeurIPS*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Jerry Liu. 2022. Llamaindex.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. 2023. Causal discovery with language models as imperfect experts. *arXiv preprint arXiv:2307.02390*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *NeurIPS*.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *J. Mach. Learn. Res.*, 17(32):1–102.

Trang Nguyen, Alexander Tong, Kanika Madan, Yoshua Bengio, and Dianbo Liu. 2023. Causal inference in gene regulatory networks with gflownet: Towards scalability in large systems. *arXiv preprint arXiv:2310.03579*.

OpenAI. 2024. Function calling guide. URL: https://platform.openai.com/docs/guides/function-calling. Accessed: 2024-12-10.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Ross Quinlan. 1993. Auto mpg. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5859H.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. 2022. Score matching enables causal discovery of nonlinear additive noise models. In *ICML*.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *NeurIPS*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *NeurIPS*.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. 2006. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7(10).

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. 2011. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *J. Mach. Learn. Res.*, 12(Apr):1225–1248.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *NeurIPS*.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet. 2017. Anomaly detection in streams with extreme value theory. In *KDD*.

Tomi Silander and Petri Myllymäki. 2006. A simple approach for finding the globally optimal bayesian network structure. In *UAI*.

David J Spiegelhalter. 1992. Learning in probabilistic expert systems. *Bayesian statistics*, 4:447–465.

Peter Spirtes and Clark Glymour. 1991. An algorithm for fast recovery of sparse causal graphs. *Soc. Sci. Comput. Rev.*, 9(1):62–72.

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

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *EMNLP*.

Ruibo Tu, Kun Zhang, Bo Bertilson, Hedvig Kjellstrom, and Cheng Zhang. 2019. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. *NeurIPS*.

Ruibo Tu, Kun Zhang, Hedvig Kjellstrom, and Cheng Zhang. 2022. Optimal transport for causal discovery. In *ICLR*.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. 2023. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*.

Dongjie Wang, Zhengzhang Chen, Yanjie Fu, Yanchi Liu, and Haifeng Chen. 2023a. Incremental causal graph learning for online root cause analysis. In *KDD*.

Dongjie Wang, Zhengzhang Chen, Jingchao Ni, Liang Tong, Zheng Wang, Yanjie Fu, and Haifeng Chen. 2023b. Interdependent causal networks for root cause localization. In *KDD*.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2024. Can language models solve graph problems in natural language? *NeurIPS*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. 2020. Generalized independent noise condition for estimating latent variable causal graphs. *NeurIPS*.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. *arXiv preprint arXiv:2305.11595*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *ICLR*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. 2025. Causaleval: Towards better causal reasoning in language models. In *NAACL-HLT*.

Changhe Yuan and Brandon Malone. 2013. Learning optimal bayesian networks: A shortest path perspective. *J. Artif. Intell. Res.*, 48:23–65.

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

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2023. Competeai: Understanding the competition behaviors in large language model-based agents. *arXiv preprint arXiv:2310.17512*.

Lecheng Zheng, Zhengzhang Chen, Dongjie Wang, Chengyuan Deng, Reon Matsuoka, and Haifeng Chen. 2024a. Lemma-rca: A large multi-modal multi-domain dataset for root cause analysis. *arXiv preprint arXiv:2406.05375*.

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

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. 2024b. Causal-learn: Causal discovery in python. *J. Mach. Learn. Res.*, 25(60):1–8.

A Algorithm

The algorithm of MATMCD is summarized in Algorithm 1. The notations are consistent with §3.

Algorithm 1: Multi-Agent with Tool-Augmented LLMs for Multi-Modality Enhanced Causal Discovery (MATMCD)

```
Input: (1) Observational Data: ${\mathbf{v}_1, ..., \mathbf{v}_n}$, where $\mathbf{v}_i = \{v_{i1}, ..., v_{im}\}$ includes $m$ samples for
       variable $v_i$; (2) Meta-Data $\mathcal{D} = \{\mathsf{s}, \mathcal{Z}\}$, where s is a descriptive title of the dataset and
       $\mathcal{Z} = \{\mathsf{z}_1, ..., \mathsf{z}_n\}$ includes the descriptive name $\mathsf{z}_i$ of each variable $v_i$ in $\mathcal{V}$; (3) An SCD
       algorithm $f_{\text{SCD}}(\cdot)$; (4) The number of guesses $K$ in Top-K Guess prompt for eliciting
       verbal confidence in Constraint LLM.
Output: Refined causal graph $\mathbf{G}$
/* Causal graph estimator (§3.1) */
Initial Causal Graph $\mathbf{G}_0 \leftarrow$ CausalGraphEstimator$(f_{\text{SCD}}, \{\mathbf{v}_1, ..., \mathbf{v}_n\})$
/* DA-agent (§3.2) */
Initialize call history memory $\mathcal{C} \leftarrow \emptyset$
Initialize retrieved data memory $\mathcal{R} \leftarrow \emptyset$
Query $\leftarrow$ Search-LLM(s, $\mathcal{Z}, \mathcal{C}$)    // Generate a search query after checking memory $\mathcal{C}$
while Query $\neq$ "No query needed" do
    /* Iterative search */
    $\mathcal{R} \leftarrow$ Toolkit(Query)                    // Add retrieved data to memory $\mathcal{R}$
    $\mathcal{C} \leftarrow$ Query                             // Add search query to memory $\mathcal{C}$
    Query $\leftarrow$ Search-LLM(s, $\mathcal{Z}, \mathcal{C}$)    // Update search query after checking memory $\mathcal{C}$
end
ContextualData $\leftarrow$ Summary-LLM(s, $\mathcal{Z}, \mathcal{R}$)
/* CC-agent (§3.3) */
Initialize constraint matrix $\mathbf{C}$
for Each pair of variables $(v_i, v_j)$ do
    Explanation$_{i,j}$ $\leftarrow$ Knowledge-LLM($\mathbf{G}_0, v_i, v_j$, ContextualData)
    $\{(\text{Constraint}_k, \text{Confidence}_k)\}_{k=1}^{K} \leftarrow$ Constraint-LLM(Explanation$_{i,j}$)  // Top-K guesses
    $k_* \leftarrow \text{argmax}_k \text{Confidence}_k$
    $\mathbf{C}_{i,j} \leftarrow \text{Constraint}_{k_*}$    // Update constraint matrix with Constraint$_{k_*}$ for $(v_i, v_j)$
end
/* Causal Graph Refiner (§3.4) */
Refined causal graph $\mathbf{G} \leftarrow$ CausalGraphRefiner$(f_{\text{SCD}}, \{\mathbf{v}_1, ..., \mathbf{v}_n\}, \mathbf{C})$
```

B Experimental Details

B.1 Implementation Details

For PC algorithm (Spirtes and Glymour, 1991), Fisher's Z test was used in its conditional independence tests. For Exact Search (Yuan and Malone, 2013), its k-cycle heuristic was enabled with $k = 2$ and its max parents was limited to 2. For DirectLiNGAM (Shimizu et al., 2011), soft prior knowledge integration was adopted to circumvent its potential conflicts with LLM generated constraints. For LLMs, we set the temperature to 0.5 by default. However, certain LLMs (*e.g.*, Llama 3.1, Mistral) may produce outputs that do not adhere to the prompt requirements. In this case, we attempt tuning the temperature to 0.7 and rerunning multiple times. For Efficient-CDLMs (Jiralerspong et al., 2024), it may produce different results across multiple runs. Thus we reported its best performance of 10 runs on each dataset.

For the implementation of RAG, `LlamaIndex` was used with `text-embedding-ada-002` for document chunk indexing. For the web search tool in DA-AGENT, Google search API via Serper[2] was adopted. If

[2] https://serper.dev/

the returned information is insufficient or incomplete, Tavily[3] was employed as a supplement. By consolidating information from multiple sources, we ensure a comprehensive and diverse context, aiding causal discovery and LLM prompts. For the log lookup tool, template-based logs were used to systematically record events. Specifically, events were capped by 10 occurrences with random sampling.

B.2 Evaluation Metrics in RCA

To evaluate the performance of RCA, following (Wang et al., 2023b; Zheng et al., 2024a), we use the following widely-used metrics.

Precision@K (PR@K) measures the probability that the Top-K predicted root cause are true, which is defined as

$$\mathrm{PR@K} = \frac{1}{|\mathbb{A}|} \sum_{a \in \mathbb{A}} \frac{\sum_{i<K} R_a(i) \in V_a}{\min(K, |V_a|)}$$

where $\mathbb{A}$ is the set of system faults, a is one fault in $\mathbb{A}$, V_a is the set of true root causes of fault a, R_a is the set of predicted root causes of fault a, and $R_a(i)$ is the i-th ranked prediction in R_a.

Mean Average Precision@K (MAP@K) assesses the model performance in the top-K predicted root causes from the overall perspective, which is defined as following.

$$\mathrm{MAP@K} = \frac{1}{K|\mathbb{A}|} \sum_{a \in \mathbb{A}} \sum_{1 \le j \le K} \mathrm{PR@j}$$

where a higher value indicates better performance.

Mean Reciprocal Rank (MRR) measures the ranking capability of models. A higher MRR indicates that the predicted root causes tend to appear earlier in the ranking. MRR is defined as following.

$$\mathrm{MRR} = \frac{1}{|\mathbb{A}|} \sum_{a \in \mathbb{A}} \frac{1}{\mathrm{rank}_{R_a}}$$

where rank_{R_a} is the rank of the first correctly predicted root cause of system fault a.

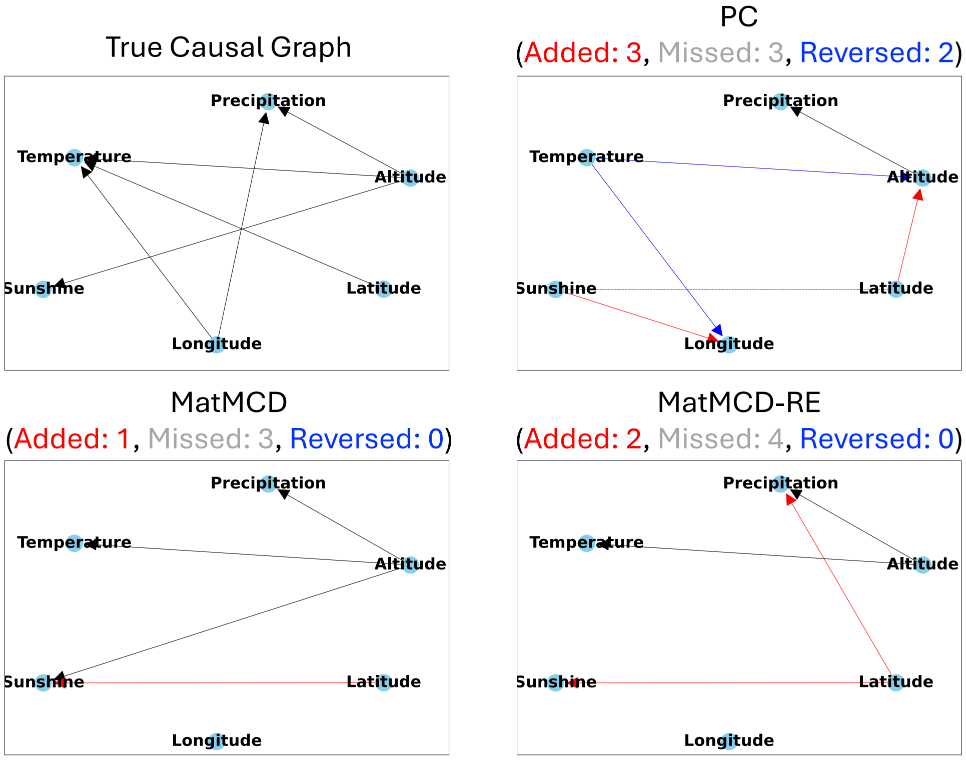
B.3 Additional Graph Visualizations

In this section, we present the visualization results of datasets `DWDClimate`, `SachsProtein`, `Asia` and `Child` in Fig. 6, Fig. 7, Fig. 8, and Fig. 9. Each figure includes the true causal graph, the causal graph predicted by the PC algorithm, the causal graph generated by MATMCD, and the causal graph generated by MATMCD-RE. In the figures, red arrows indicate edges that were added by error, blue arrows denote edges that were reversed by error. In each figure, we also included the number of wrongly Added edges, Missed edges, and Reversed edges for comparison.

Compared to the PC algorithm, MATMCD(-RE) yields causal graphs that resemble the true causal graph better, with in general less numbers of wrongly Added, Missed, and Reversed edges. This advantage stems from the LLM agents' ability to impose multi-modal data empowered knowledge-driven constraints for screening erroneous edges. In contrast, the PC algorithm only relies on correlations of the observational data between variables, rendering it more error-prone.

Furthermore, unlike the PC algorithm, which relies on an undirected graph structure to determine the direction of causal edges, the incorporation of LLM-provided knowledge offers precise guidance in establishing causal directions. Therefore, the proposed MATMCD(-RE) exhibits fewer reversed edges on most of the datasets, particularly on the datasets with more accessible multi-modal data (*e.g.*, the non-biomedical datasets).

[3] `https://tavily.com/`

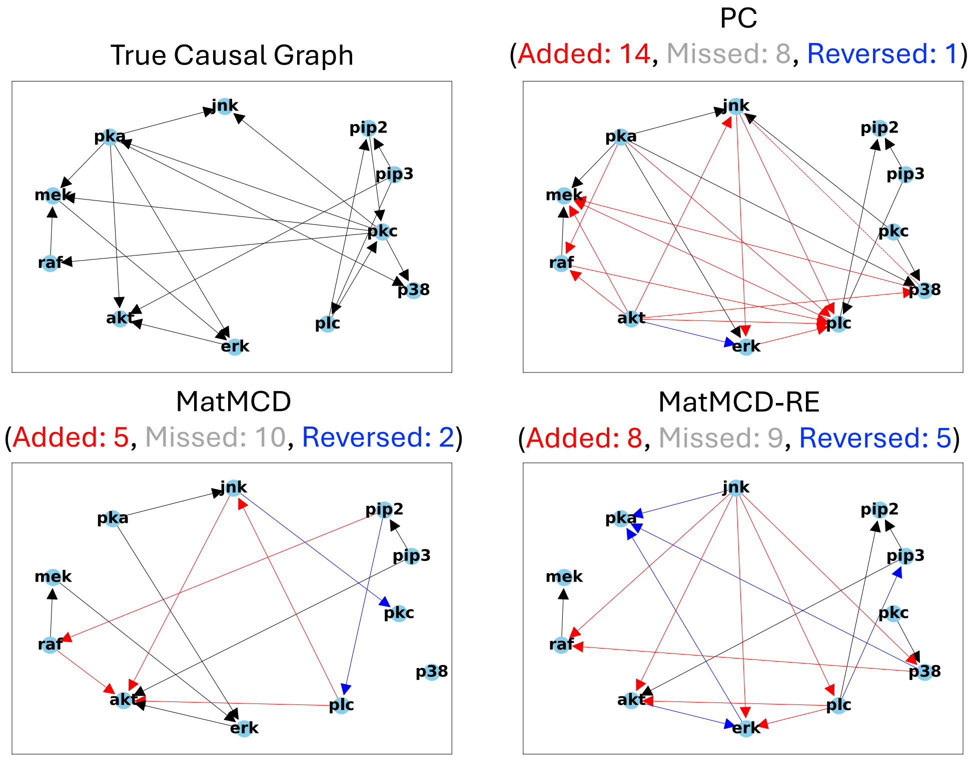

True Causal Graph

PC (Added: 3, Missed: 3, Reversed: 2)

MatMCD (Added: 1, Missed: 3, Reversed: 0)

MatMCD-RE (Added: 2, Missed: 4, Reversed: 0)

Figure 6: Causal Graph visualization for `DWDClimate` dataset.

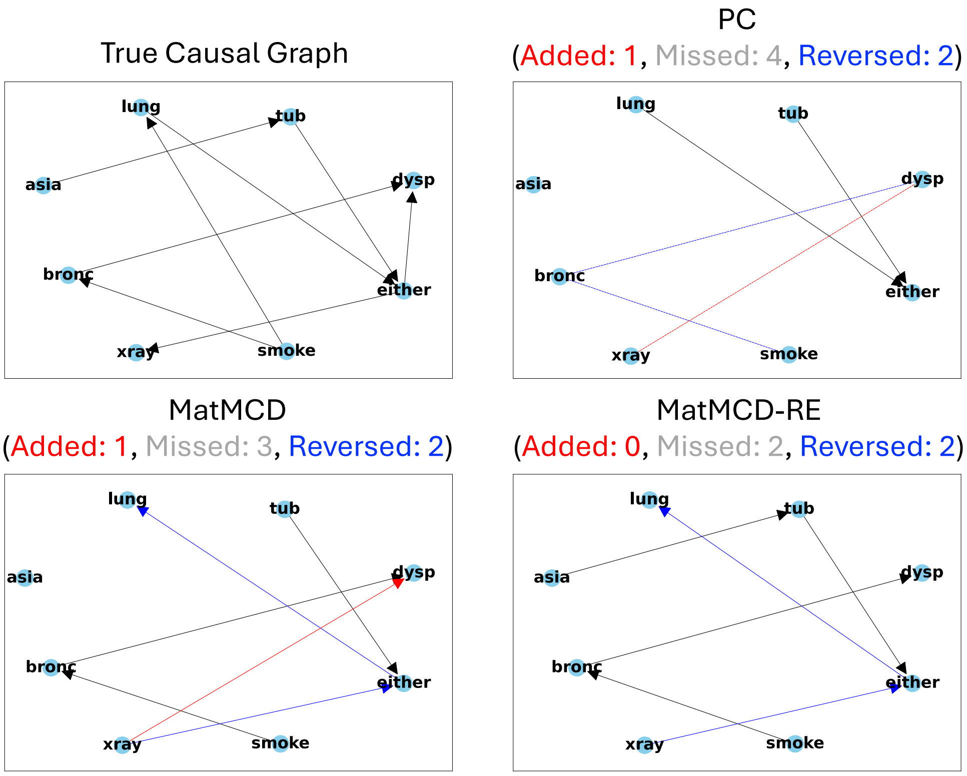

True Causal Graph

PC (Added: 14, Missed: 8, Reversed: 1)

MatMCD (Added: 5, Missed: 10, Reversed: 2)

MatMCD-RE (Added: 8, Missed: 9, Reversed: 5)

Figure 7: Causal Graph visualization for `SachsProtein` dataset.

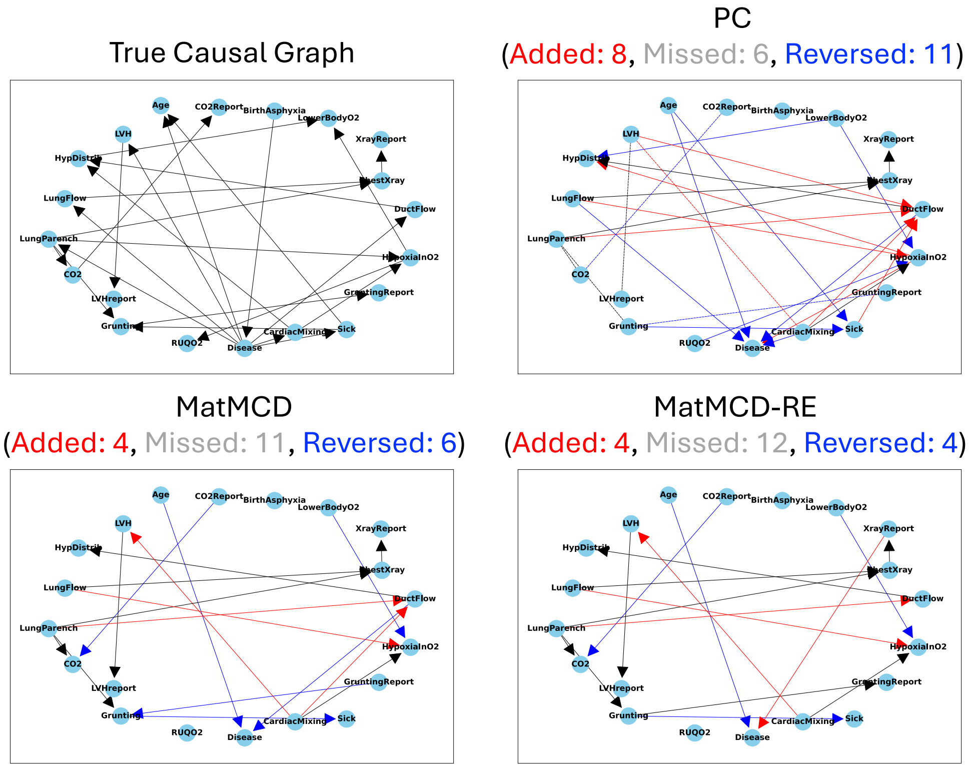

True Causal Graph

PC
(Added: 1, Missed: 4, Reversed: 2)

MatMCD
(Added: 1, Missed: 3, Reversed: 2)

MatMCD-RE
(Added: 0, Missed: 2, Reversed: 2)

Figure 8: Causal Graph visualization for Asia dataset.


True Causal Graph

PC
(Added: 8, Missed: 6, Reversed: 11)

MatMCD
(Added: 4, Missed: 11, Reversed: 6)

MatMCD-RE
(Added: 4, Missed: 12, Reversed: 4)

Figure 9: Causal Graph visualization for Child dataset.

C Prompt Templates

In this section, we provide the prompt templates for the Search LLM and Summary LLM in DA-AGENT (§3.2), the Knowledge LLM and Constraint LLM in CC-AGENT (§3.3). Additionally, we include the prompt template for the single-round search used in the Ablation Analysis in Table 3(a) and the prompt template for the CC-AGENT without Knowledge LLM used in Ablation Analysis in Table 3(b).

C.1 Data Augmentation Agent (DA-AGENT)

Search LLM

```
Assume you have no prior knowledge about the <dataset name> dataset and its nodes <list of node names>. You need to generate queries for others who can search for information that will help you summarize the dataset and the nodes, and help you recognize the relationships between the nodes.

Split your query into multiple sub-queries. Ensure each query is specific, clear, and easy to be used for search. If you have inquired about this topic before, your previous queries are:

<list of previous queries>

Please try to avoid repeating these queries.  Generate a new query to get more information. Format your query as follows:

Search Query: <new query>

You should generate only one query at a time. If you think no additional queries are needed, please state ‘No query needed’.
```

Summary LLM (with RAG)

```
Please provide a summary of the <dataset name> dataset and its nodes including <list of node names> using the information from the RAG Database.  Your response should include detailed information on the dataset and each of its nodes.

Follow the structured format below in your answer:

- Dataset Summary: <a general summary of the dataset>

- Summary of <node name>: <a detailed summary of this node>

- Summary of <node name>: <a detailed summary of this node>

Additionally, try to include information on the relationships between the nodes after you have summarized each one. Make sure to cover all the nodes mentioned.
```

Summary LLM (with Log)

In <dataset name> dataset, the following entities are included: <list of node names>. The log information about the entity <node name> is provided in the following format:

<log format template>

Below are the information in the log following the above format:

<The list of node's event templates and examples>

Based on the above information, please provide your summary of this entity: <node name>, including:

- The role of the entity: <node name> in the system

- The key events that need to be noticed. Need to pay attention to both frequent and infrequent occurrences

- The status of this entity

- The relationships this entity: <node name> has with other entities in the above entity list. Only include the most likely ones.

You only need to provide the information about this entity: <node name>. Your response should be in the following format:

The name of the entity: <node name>

Role of the entity: <role>

Key events that need to be noticed: <key events>

The status of this entity: <status>

The relationships of <node name>: <relationships>

Your response:

C.2 Causal Constraint Agent (CC-AGENT)

Knowledge LLM

We want to perform causal discovery on <dataset name> , the summary of dataset: <dataset information>. Considering <list of node names> as variables. We have conducted the statistical causal discovery with <SCD algorithm name> algorithm.

The edges and their coefficients of the causal structure suggested by the statistical discovery are as follows:

<adjacency list of the causal graph>

Based on the information above, it seems that changes in <node name i> have <a/no> direct impact on <node name j>. In addition, here is the information of <node name i> and <node name j> from reliable sources:

<information about node i>

<information about node j>

Your task is to interpret this result from a domain knowledge perspective and determine whether this statistically suggested hypothesis is plausible in the context of the domain. Please provide an explanation that leverages your expert knowledge on the causal relationship between <node name i> and <node name j>, and assess the correctness of this causal discovery result.

Your response should consider the relevant factors and provide a reasonable explanation based on your understanding of the domain.

Constraint LLM

Provide your <K> best guesses and the probability that each is correct (0.0 to 1.0) for the following question. Give ONLY the guesses and probabilities, no other words or explanation. Each guess should infer the relationship step by step and finally end with <Yes> or <No>. For example:

G1: <the first most likely guess, infer the relationship step by step and end with <Yes> or <No> >

P1: <the probability between 0.0 and 1.0 that G1 is correct, without any extra comments; just the probability!>

G2: <the second most likely guess, infer the relationship step by step and end with <Yes> or <No> >

P2: <the probability between 0.0 and 1.0 that G2 is correct, without any extra comments; just the probability!>

The question is: here is the explanation from an expert in the field of <dataset name> regarding the causal relationship between <node name i> and <node name j>:

<domain knowledge output of Knowledge LLM>

Considering the information above, if <node name i> is modified, will it have a direct impact on <node name j>?

C.3 Single-Round Search

In the implementation of single-round search, SearchGPT[4] was used in combination with the prompt below to directly search for pertinent information about the dataset and nodes, thus the LLM is not required to generate a query.

Single-Round Search

Please provide detailed information of dataset <dataset name> and its variables, including <list of node names>. Try to search for more information if possible. You should provide information about each variable and their relationships with each other. Please provide detailed information.

Follow the structured format below in your answer:

The summary of dataset <dataset name>:

- Dataset Summary: <a general summary of the dataset>

- Summary of <node name>: <a detailed summary of this node>

- Summary of <node name>: <a detailed summary of this node>

Additionally, try to include information on the relationships between the nodes after you have summarized each one. Make sure to cover all the nodes mentioned.

C.4 CC-AGENT without Knowledge LLM

CC-AGENT without Knowledge LLM

We want to perform causal discovery on <dataset name>, the summary of dataset: <dataset information>. Considering <list of node names> as variables. We have conducted the statistical causal discovery with <SCD algorithm name> algorithm.

The edges and their coefficients of the causal structure suggested by the statistical discovery are as follows:

<adjacency list of the causal graph>

Based on the information above, it seems that changes in <node name i> have <a/no> direct impact on <node name j>. In addition, here is the information of <node name i> and <node name j> from reliable sources:

<information about node i>

<information about node j>

Do you think that a change in <node name i> will lead to a change in <node name j>, which means <node name i> has a causal influence on <node name j>?

Please answer this question with ⟨Yes⟩ or ⟨No⟩. No other responses are needed. Your response should be in the following format: ⟨Yes⟩ or ⟨No⟩.

[4] https://github.com/Wilson-ZheLin/SearchGPT

D Examples of LLM Response

D.1 Data Augmentation Agent (DA-AGENT)

In this section, we provide the response of LLMs during the search process of DA-AGENT on the AutoMPG dataset.

Query 1 (from Search LLM): what is the auto mpg dataset, and what are the variables displacement, mpg, horsepower, weight, and acceleration?

Answer 1: The auto mpg dataset is a well-known dataset in the field of machine learning and statistics, primarily used for regression analysis and predictive modeling. it contains information about various automobiles, specifically focusing on their fuel efficiency measured in miles per gallon (mpg). the dataset includes several variables that are crucial for understanding vehicle performance, including displacement, mpg, horsepower, weight, and acceleration.

1. **displacement**: this variable refers to the total volume of all the cylinders in an engine, typically measured in cubic inches or liters. it is a key indicator of engine size and capacity. generally, a larger displacement can lead to more power output, but it may also result in lower fuel efficiency.
2. **mpg (miles per gallon)**: this is the target variable in the dataset, representing the fuel efficiency of the vehicle. higher mpg values indicate better fuel efficiency, meaning the vehicle can travel further on a gallon of fuel. the relationship between mpg and other variables is often inverse; as engine size (displacement) and weight increase, mpg tends to decrease.
3. **horsepower**: this variable measures the power output of the engine. higher horsepower typically correlates with better acceleration and overall performance. however, like displacement, higher horsepower can also lead to lower fuel efficiency, as more power often requires more fuel.
4. **weight**: the weight of the vehicle is another critical factor affecting performance. heavier vehicles generally require more power to accelerate, which can negatively impact mpg. there is often a negative correlation between weight and mpg; as weight increases, fuel efficiency tends to decrease.
5. **acceleration**: this variable measures how quickly a vehicle can increase its speed, usually defined as the time taken to go from 0 to 60 mph. Acceleration is influenced by both horsepower and weight; more powerful engines can accelerate faster, while heavier vehicles may struggle to achieve the same acceleration.

In summary, the relationships among these variables can be summarized as follows:

- **Displacement and Horsepower**: Generally, larger displacement engines produce more horsepower.
- **Horsepower and Acceleration**: Higher horsepower typically results in better acceleration.
- **Weight and MPG**: Increased weight usually leads to lower MPG.
- **Displacement and MPG**: Larger engines (higher displacement) often result in lower MPG due to increased fuel consumption.

Understanding these relationships is crucial for analyzing vehicle performance and making informed decisions regarding vehicle design and efficiency improvements.

Query 2 (from Search LLM): What are the relationships between its variables such as Displacement, MPG, Horsepower, Weight, and Acceleration?

Answer 2: The Auto MPG dataset is a well-known dataset that contains information about various attributes of automobiles and their fuel consumption, specifically measured in miles per gallon (MPG). The dataset includes five key continuous variables: **Weight**, **Displacement**, **Horsepower**, **Acceleration**, and **MPG**. It is often used in data analysis and machine learning to understand the factors that influence fuel efficiency in cars.

Relationships Between Variables

1. **Displacement**: Displacement refers to the engine size of the car, and it has been found to have a strong negative correlation with MPG. This means that as the displacement increases, the MPG tends to decrease. This relationship is intuitive, as larger engines typically consume more fuel.
2. **Horsepower**: Similar to displacement, horsepower also shows a negative causal effect on MPG. Higher horsepower engines are generally less fuel-efficient, which is supported by statistical analysis indicating a significant negative relationship between horsepower and MPG.
3. **Weight**: The weight of the car is another critical factor affecting fuel efficiency. Heavier cars tend to have lower MPG values, as more energy is required to move a heavier vehicle. This relationship is also confirmed through causal inference results, indicating that increasing the weight of a car significantly reduces its fuel efficiency.
4. **Acceleration**: Acceleration has a complex relationship with MPG. While faster acceleration can be associated with lower fuel efficiency, the analysis suggests that the relationship is not as straightforward as with displacement and horsepower. The causal effect of acceleration on MPG is negative, indicating that cars designed for quicker acceleration may sacrifice fuel efficiency.
5. **MPG**: MPG is a key indicator for measuring vehicle fuel efficiency, reflecting the distance a vehicle can travel on a certain amount of fuel. This indicator is influenced by various factors, including weight, displacement, horsepower, and acceleration.

Summary of Findings

- **Negative Correlations**: Displacement, horsepower, and weight are negatively correlated with MPG, meaning that increases in these variables lead to decreases in fuel efficiency.
- **Complex Relationships**: Acceleration has a negative effect on MPG, but the relationship is influenced by other factors and may not be as direct as with weight or displacement.

In conclusion, the Auto MPG dataset provides valuable insights into how various automobile attributes affect fuel consumption. Understanding these relationships can help in designing more fuel-efficient vehicles and improving overall automotive performance.

Example Summary of `AutoMPG` Dataset

Dataset Summary: The Auto MPG dataset is a well-known dataset in the field of machine learning and statistics, primarily used for regression analysis and predictive modeling. It contains information about various automobiles, specifically focusing on their fuel efficiency measured in miles per gallon (MPG). The dataset includes several variables that are crucial for understanding vehicle performance, including Displacement, Mpg, Horsepower, Weight, and Acceleration.

`/* Each of the following summaries describes a variable and its relationships with other variables */`

`---`

Summary of Displacement: Displacement refers to the total volume of all the cylinders in an engine, typically measured in cubic inches or liters. It is a key indicator of engine size and capacity. Generally, a larger displacement can lead to more power output, but it may also result in lower fuel efficiency.

`---`

Summary of Mpg: Mpg (Miles per Gallon) is the target variable in the dataset, representing the fuel efficiency of the vehicle. Higher MPG values indicate better fuel efficiency, meaning the vehicle can travel further on a gallon of fuel. The relationship between MPG and other variables is often inverse; as engine size (Displacement) and weight increase, MPG tends to decrease.

`---`

Summary of Horsepower: Horsepower measures the power output of the engine. Higher horsepower typically correlates with better acceleration and overall performance. However, like displacement, higher horsepower can also lead to lower fuel efficiency, as more power often requires more fuel.

`---`

Summary of Weight: The weight of the vehicle is another critical factor affecting performance. Heavier vehicles generally require more power to accelerate, which can negatively impact MPG. There is often a negative correlation between weight and MPG; as weight increases, fuel efficiency tends to decrease.

`---`

Summary of Acceleration: Acceleration measures how quickly a vehicle can increase its speed. It is influenced by both horsepower and weight; more powerful engines can accelerate faster, while heavier vehicles may struggle to achieve the same acceleration.

D.2 Causal Constraint Agent (CC-AGENT)

In this section, we provide two examples of knowledge generated by the Knowledge LLM in CC-AGENT, including cases where the Knowledge LLM agrees with and opposes the results of the SCD algorithm.

Domain Knowledge from Displacement → MPG

The findings from the statistical causal discovery using the PC algorithm suggest a complex interplay between the variables in the AutoMPG dataset, particularly highlighting the relationships between Displacement, Horsepower, Weight, Acceleration, and Mpg. To interpret these results from a domain knowledge perspective, let's break down the relationships and their implications for gasoline consumption.

Understanding the Variables

- **Displacement**: This is a measure of the engine's size and capacity. Generally, larger displacement engines can produce more power, which can be beneficial for performance. However, larger engines often consume more fuel, leading to lower fuel efficiency (MPG).
- **Mpg (Miles per Gallon)**: This is the key measure of fuel efficiency. Higher MPG values indicate that a vehicle can travel further on a gallon of fuel, which is desirable for cost savings and environmental considerations.

Interpretation of the Findings

The conclusion that changes in Displacement do not directly affect Mpg is plausible and aligns with our understanding of automotive engineering and fuel efficiency. Here are some key considerations:

1. **Indirect Relationships**: While Displacement itself does not directly influence Mpg, it affects other variables (Horsepower, Weight, Acceleration) that subsequently impact MPG. This indicates a multi-layered causal structure where Displacement plays a foundational role, but its effects are mediated through other factors.
2. **Complex Interactions**: The relationships among these variables are not straightforward. For example, a vehicle with a larger engine might be heavier and more powerful, but if it is designed with fuel efficiency in mind (e.g., hybrid technology), it could achieve a higher MPG despite the larger displacement.
3. **Empirical Evidence**: Empirical studies in automotive engineering support the notion that while larger engines tend to be less fuel-efficient, the design choices and technologies implemented in vehicles can significantly alter this relationship. For instance, advancements in engine technology (like turbocharging) can allow for smaller engines to produce more power while maintaining or improving fuel efficiency.

Conclusion

The causal discovery results are consistent with domain knowledge. Displacement does not directly affect MPG, but it influences other factors that collectively determine fuel efficiency. This nuanced understanding reflects the complexity of automotive design and performance, where multiple variables interact in ways that may not be immediately apparent.

Overall, the findings emphasize the importance of considering the entire system of vehicle characteristics rather than isolating single variables when assessing gasoline consumption and fuel efficiency. This insight is crucial for developing more efficient vehicles and informing consumer choices.

Domain Knowledge from MPG → Weight

Causal Relationships Overview

In analyzing causal relationships within the Auto MPG dataset, we expect to see significant impacts from vehicle characteristics like engine displacement, horsepower, and weight on MPG. However, the idea that MPG could directly influence a vehicle's weight is counterintuitive based on principles of vehicle design and physics.

Assessment of the Claim "MPG Influences Weight"

1.**Physical and Engineering Principles**: In conventional vehicle engineering, weight is primarily determined by the vehicle's physical construction, such as body materials, engine size, and overall design. MPG is an outcome metric that reflects the fuel efficiency of a vehicle under specific conditions, not a factor that could in turn influence these physical characteristics.

2.**Design Prioritization**: Vehicle designers prioritize meeting performance and safety standards before optimizing for fuel efficiency. This means that decisions about weight are made early in the design and manufacturing process, and MPG is a derivative outcome after these factors are considered. Thus, it is logically inconsistent to view MPG as a direct influencer of weight.

3.**Difference Between Statistical Correlation and Causation**: Even if statistical analysis shows a correlation between MPG and weight, this relationship should not be misconstrued as causative. Improvements in MPG are often achieved through the use of lighter materials and more efficient engines, but these improvements primarily affect the design and manufacturing process, not MPG directly acting on weight.

Reevaluating the Role of MPG

Indirect Influence: Although there is an association between MPG and a vehicle's weight, such as lighter vehicles often having higher MPG, this relationship is more about weight influencing MPG rather than MPG influencing weight. The use of lightweight materials and efficient technologies is primarily aimed at achieving better performance and environmental standards, which may indirectly improve MPG.

Conclusion

Based on fundamental principles of vehicle manufacturing and design, MPG does not directly influence a vehicle's weight. MPG results from the interaction of various factors, not as a cause of changes in those factors. In analyzing automotive data, it is crucial to distinguish between outcomes and causes and understand how these variables interact in real engineering and design contexts. Therefore, suggesting that MPG influences vehicle weight is a misconception, reflecting a misunderstanding of the vehicle design process and causal relationships.