# RealTCD: Temporal Causal Discovery from Interventional Data with Large Language Model

Peiwen Li*
SIGS, Tsinghua University
lpw22@mails.tsinghua.edu.cn

Xin Wang†
DCST, BNRist, Tsinghua University
xin_wang@tsinghua.edu.cn

Zeyang Zhang
DCST, Tsinghua University
zy-zhang20@mails.tsinghua.edu.cn

Yuan Meng†
DCST, Tsinghua University
yuanmeng@tsinghua.edu.cn

Fang Shen
Alibaba Cloud
ziru.sf@alibaba-inc.com

Yue Li
Alibaba Cloud
yueqian.ly@alibaba-inc.com

Jialong Wang†
Alibaba Cloud
quming.wjl@alibaba-inc.com

Yang Li
SIGS, Tsinghua University
yangli@sz.tsinghua.edu.cn

Wenwu Zhu†
DCST, BNRist, Tsinghua University
wwzhu@tsinghua.edu.cn

## Abstract

In the field of Artificial Intelligence for Information Technology Operations, causal discovery is pivotal for operation and maintenance of systems, facilitating downstream industrial tasks such as root cause analysis. Temporal causal discovery, as an emerging method, aims to identify temporal causal relations between variables directly from observations by utilizing interventional data. However, existing methods mainly focus on synthetic datasets with heavy reliance on interventional targets and ignore the textual information hidden in real-world systems, failing to conduct causal discovery for real industrial scenarios. To tackle this problem, in this paper we investigate temporal causal discovery in industrial scenarios, which faces two critical challenges: how to discover causal relations without the interventional targets that are costly to obtain in practice, and how to discover causal relations via leveraging the textual information in systems which can be complex yet abundant in industrial contexts. To address these challenges, we propose the **RealTCD** framework, which is able to leverage domain knowledge to discover temporal causal relations without interventional targets. We first develop a score-based temporal causal discovery method capable of discovering causal relations without relying on interventional targets through strategic masking and regularization. Then, by employing Large Language Models (LLMs) to handle texts and integrate domain knowledge, we introduce LLM-guided meta-initialization to extract the meta-knowledge from textual information hidden in systems to boost the quality of discovery. We conduct extensive experiments on both simulation datasets and our real-world application scenario to show the superiority of our proposed **RealTCD** over existing baselines in temporal causal discovery.

## CCS Concepts

• **Computing methodologies** → *Causal reasoning and diagnostics*; *Temporal reasoning*.

## Keywords

Large Language Model, Causal Discovery, Time Series, Intervention

**ACM Reference Format:**
Peiwen Li, Xin Wang, Zeyang Zhang, Yuan Meng, Fang Shen, Yue Li, Jialong Wang, Yang Li, and Wenwu Zhu. 2024. RealTCD: Temporal Causal Discovery from Interventional Data with Large Language Model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3627673.3680042

## 1 Introduction

The advent of Artificial Intelligence for Information Technology Operations (AIOps) has revolutionized the way we manage and operate complex information systems. Causal discovery plays a pivotal role in understanding the intricate network of dependencies and influences within these systems [2, 30, 57, 59], offering invaluable insights for various downstream industrial tasks in AIOps, including anomaly detection [54] and root cause analysis [45, 53] etc. For instance, by equipping AIOps with the ability to accurately identify the underlying causal structures, AIOps systems can effectively detect abnormal behaviors and determine the underlying causes of system failures, thus leading to enhanced operational efficiency and improved decision-making processes in the industry.

Temporal causal discovery, as an emerging approach, aims to directly identify temporal causal relationships between variables based on observational data, with the utilization of interventional data. This group of methods has gained significant attention in recent years due to their promising potential to uncover causal dependencies in dynamic systems. Brouillard et al. [6] and Li et al. [26] employ temporal causal discovery methods to leverage various types of interventional data and have achieved remarkable progress in discovering the underlying temporal causal relationships.

However, the existing studies mainly focus on studying synthetic datasets, which strongly rely on interventional targets and ignore the intricate complexities and nuances hidden in real-world systems,

*The work was done during author's internship at Alibaba Cloud.
†Corresponding authors.

*CIKM '24, October 21–25, 2024, Boise, ID, USA*
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3680042

failing to conduct causal discovery for real industrial scenarios. In this paper, we tackle this problem by studying temporal causal discovery in industrial scenarios, which is non-trivial and poses the following two critical challenges:

- How to discover causal relationships without the interventional targets that are normally costly to obtain in practice?
- How to discover causal relationships via leveraging the textual information in systems which can be complex yet abundant in industrial contexts?

To address these challenges, we propose the **RealTCD** framework, which is able to leverage the textual information from real-world systems to discover temporal causal relationships without interventional targets. Specifically, we first develop a score-based temporal causal discovery method that learns the underlying causal relationship without interventional targets through strategic masking and regularization. We impose regularizations on both the adjacency matrix and interventional family within the context of the regularized maximum log-likelihood score, and optimize them in a joint manner. In this way, the costly interventional targets are not required for broader applications in real-world industrial scenarios. Subsequently, by leveraging Large Language Models (LLMs) to handle texts, we introduce LLM-guided meta-initialization that infers and initializes the inherent causal structures from the textual information in systems for the aforementioned discovery process, which incorporates the domain knowledge while upholding the theoretical integrity of temporal causal discovery. Extensive experiments on both simulation and real-world datasets demonstrate the superiority of our **RealTCD** framework over existing baselines. Deeper analyses also show that our method can effectively discover the underlying temporal causal relationships without interventional targets in industrial scenarios. In summary, our main contributions are as follows:

- We study the problem of temporal causal discovery in industrial scenarios. To the best of our knowledge, we are the first to solve the problem with Large Language Models (LLMs) and without interventional targets.
- We propose the **RealTCD** framework, including two specially designed modules: score-based temporal causal discovery and LLM-guided meta-initialization, which is able to leverage the textual information in systems to discover temporal causal relationships without interventional targets in industrial scenarios.
- Extensive experiments on both simulation and real-world datasets demonstrate the superiority of our framework over several baselines in discovering temporal causal structures without interventional targets.

## 2 Preliminary

### 2.1 Dynamic Causal Graphical Model

Since causal graphical models (CGMs) support interventions compared with standard Bayesian Networks, we introduce the dynamic causal graphical models (DyCGMs) extended from CGMs in order to formulate interventions between variables across time slices. Suppose that there are $d$ different measuring points in a system, and we consider causality within $p$ time-lagged terms. Therefore, the object of our study on temporal causal discovery is actually $(p+1)\times d$ random variables $N_{0,1},\ldots,N_{0,d},\ldots,N_{p,1},\ldots,N_{p,d}$, where $N_{k,l}, k \in \{0,\ldots,p\}, l \in \{1,\ldots,d\}$ denotes the $k$ time-lagged version of the $l$th measuring point. For the convenience of subsequent presentation, we abbreviate the above variables in order as $X_i, i \in \{1,\ldots,(p+1)d\}$.

Based on this, a DyCGM is defined by the distribution $P_X$ over the vector $X = (X_1,\ldots,X_{(p+1)d})$ and a DAG $\mathcal{G} = (V, E)$. To be specific, each node $i \in V = \{1,\ldots,(p+1)d\}$ is related with a random variable $X_i, i \in \{1,\ldots,(p+1)d\}$, and each edge $(i,j) \in E$ represents a direct causal relation from variable $X_i$ to $X_j$. Under the Markov assumption of the distribution $P_Y$ and graph $\mathcal{G}$, the joint distribution can be factorized as $p(x_1,\ldots,x_{(p+1)d}) = \prod_{j=1}^{(p+1)d} p_j(x_j|x_{\pi_j^{\mathcal{G}}})$ , where $\pi_j^{\mathcal{G}}$ is the set of parents of the node $j$ in the graph $\mathcal{G}$, and $x_{\pi_j^{\mathcal{G}}}$ denotes the entries of the vector $x$ with indices in $\pi_j^{\mathcal{G}}$.

We also assume *causal sufficiency*, which means there is no hidden common cause that is causing more than one variable in $X$ [35].

### 2.2 Intervention

An intervention on a variable $x_j$ corresponds to replacing its conditional density $p_j(x_j|x_{\pi_j^{\mathcal{G}}})$ by a new one. Apart from that, we define the ***interventional target***, a set $I \subseteq V$ consisting of the variables being intervened simultaneously, and the ***interventional family*** $\mathcal{I} := (I_1,\ldots,I_Q)$, where Q is the number of interventions. To be specific, the observational setting, where no variables were intervened, is always known and denoted by $I_1 := \emptyset$. The $q$th interventional joint density can be represented as

$$p^{(q)}(x_1,\ldots,x_{(p+1)d}) := \prod_{j\notin I_q} p_j^{(1)}(x_j|x_{\pi_j^{\mathcal{G}}}) \prod_{j\in I_q} p_j^{(q)}(x_j|x_{\pi_j^{\mathcal{G}}}). \qquad (1)$$

Note that, in the temporal domain, merely the contemporary variables $N_{0,1},\ldots,N_{0,d}$, i.e. $X_1,\ldots,X_d$ can be intervened and be in an interventional target, as only time-lagged variables $N_{k,l}, k \in \{1,\ldots,p\}, l \in \{1,\ldots,d\}$, i.e. $X_i, i \in \{d+1,\ldots,(p+1)d\}$ and other contemporary variables $N_{0,l}, l \in \{1,\ldots,d\}\backslash\{j\}$, i.e. $X_i, i \in \{0,\ldots,d\}\backslash\{j\}$ can affect a particular contemporary variable $N_{0,j}$, i.e. $X_j$.

Meanwhile, there are two types of interventions: 1) imperfect (or soft, parametric) interventions is the general type depicted below, and 2) perfect interventions (or hard, structural) [18] is a special case that removes the dependencies of an intervened node $j \in \mathcal{I}_q$ on its parent nodes, i.e. $p_j^{(q)}(x_j|x_{\pi_j^{\mathcal{G}}}) = p_j^{(q)}(x_j)$ in equation (1).

## 3 Method

In order to make the causal discovery process able to be applied to the real industrial scene effectively, we propose the **RealTCD** framework as shown in Figure 1, including two modules: 1) in the module **Score-based Temporal Causal Discovery**, data from normal state and abnormal state of a system under AIOps are modeled as observational data and interventional data respectively, and relaxation is done to the condition that the label of interventional targets is known, making the algorithm easier to apply to real scenes; 2) in the module **LLM-guided Meta Initialization**, LLM is leveraged to introduce the domain knowledge and system structure information in text types and to preliminarily obtain possible causal relations from them as initialization for the discovery process.

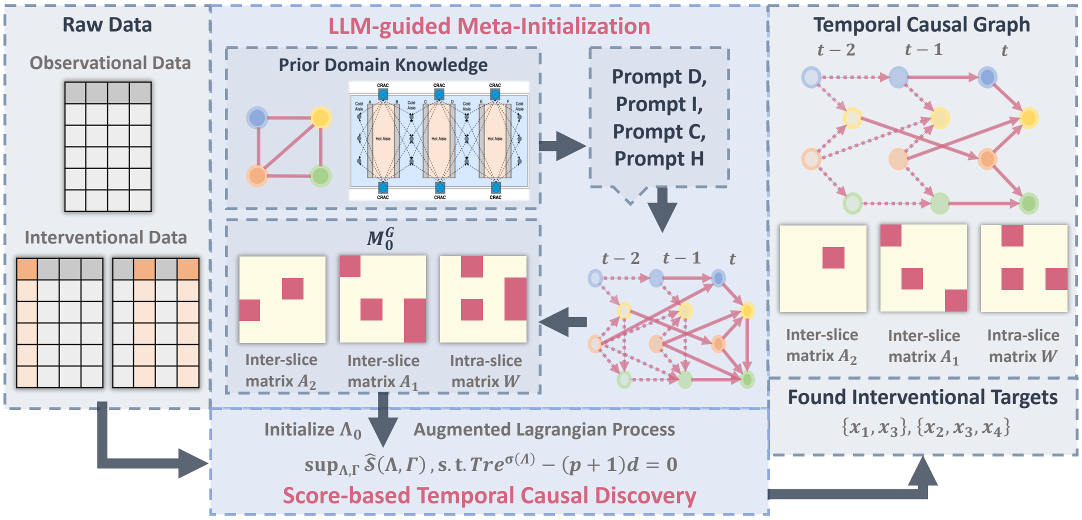


**Figure 1: The framework of our proposed method RealTCD. Given the system textual information and temporal data without interventional targets, the LLM-guided Meta-Initialization module leverages LLMs to extract the domain knowledge and obtain the potential causal relationships as the initialization adjacency matrix $M_0^{\mathcal{G}}$. Then, the Score-based Temporal Causal Discovery module utilizes an augmented Lagrangian process to optimize the score for unknown interventional targets under constraints, where the $\Lambda_0$ is initialized with $M_0^{\mathcal{G}}$. In this way, the proposed RealTCD leverages the system's textual information to discover temporal causal relationships without interventional targets.**

## 3.1 Score-based Temporal Causal Discovery

In this section, we introduce a score-based purely data-driven temporal causal discovery from interventional data with unknown interventional targets.

*3.1.1 Raw data.* Raw data we used as shown in Figure 1 refers to a set of input temporal data including both observational and interventional data. Observational data refers to standard data without interventions or anomalies, while interventional data contains interventions discussed in Section 2.2, or represents abnormal data in real-world datasets. The colored columns represent time series of ground truth interventional targets that are unknown in prior. There may be multiple sets of interventional data, each corresponding to different interventional targets or anomalies.

*3.1.2 Model conditional densities.* To begin with, we use neural networks to model conditional densities. Firstly, we encode the DAG $\mathcal{G}$ with a binary adjacency matrix $M^{\mathcal{G}} \in \{0,1\}^{(p+1)d \times (p+1)d}$ which acts as a mask on the neural network inputs. Similarly, we encode the interventional family $\mathcal{I}$ with a binary matrix $R^{\mathcal{I}} \in \{0,1\}^{Q \times (p+1)d}$, where $R^{\mathcal{I}}_{qj} = 1$ means that $X_j$ is an intervened node in the interventional target set $I_q$. Then, following equation (1), we further model the joint density of the $q$th intervention by

$$f^{(q)}\left(x; M^{\mathcal{G}}, R^{\mathcal{I}}, \phi\right) := \prod_{j=1}^{(p+1)d} \tilde{f}\left(x_j; \mathrm{NN}\left(M_j^{\mathcal{G}} \odot x; \phi_j^{(1)}\right)\right)^{1-R^{\mathcal{I}}_{qj}} \tilde{f}\left(x_j; \mathrm{NN}\left(M_j^{\mathcal{G}} \odot x; \phi_j^{(q)}\right)\right)^{R^{\mathcal{I}}_{qj}}, \tag{2}$$

where $\phi := \{\phi^{(1)}, \ldots, \phi^{(Q)}\}$, the NN's are neural networks parameterized by $\phi_j^{(1)}$ or $\phi_j^{(q)}$, the operator $\odot$ denotes the Hadamard product (element-wise) and $M_j^{\mathcal{G}}$ denotes the $j$th column of $M^{\mathcal{G}}$, enabling $M_j^{\mathcal{G}} \odot x$ to select the parents of node $j$ in the graph $\mathcal{G}$.

*3.1.3 Score for unknown interventional targets.* Based on the NN conditional densities in equation (2), we can firstly formulate the following **regularized maximum log-likelihood score** as the basic score for known interventional targets' setting:

$$\mathcal{S}_{\mathcal{I}^*}(\mathcal{G}) := \sup_{\phi} \sum_{q=1}^{Q} \mathbb{E}_{X \sim p^{(q)}} \log f^{(q)}\left(X, M^{\mathcal{G}}, R^{\mathcal{I}^*}, \phi\right) - \lambda|\mathcal{G}|, \tag{3}$$

where the ground truth interventional family (containing the interventional targets) $\mathcal{I}^* := (I_1^*, \ldots, I_Q^*)$ is known and $p^{(q)}$ stands for the $q$th ground truth interventional distribution, $|\mathcal{G}|$ represents the number of edges in the causal graph. By maximizing the score in equation (3), we can get an estimated DAG $\hat{\mathcal{G}}$ that is $\mathcal{I}^*$-Markov equivalent to the true DAG $\mathcal{G}^*$ [6], under the condition that the ground truth interventional family is known.

Then, we assume the interventional targets are unknown. To still be able to utilize the special information from interventional data, we propose to jointly optimize the adjacency matrix and interventional family as well as the NN's parameters, thus, reaching the two optimization goals simultaneously. To be specific, a regularization term for the interventional family is added to the above score, and we form score for unknown interventional targets:

$$\mathcal{S}(\mathcal{G}, \mathcal{I}) := \sup_{\phi} \sum_{q=1}^{Q} \mathbb{E}_{X \sim p^{(q)}} \log f^{(q)}\left(X, M^{\mathcal{G}}, R^{\mathcal{I}}, \phi\right) - \lambda|\mathcal{G}| - \lambda_R|\mathcal{I}|, \tag{4}$$

where $|\mathcal{I}| = \sum_{q=1}^{Q} |\mathcal{I}_q|$ counts the total number of intervened nodes.

The following theorem guarantees the identification of the temporal causal graph as well as the interventional family under the setting that interventional targets are unknown, which can be proved similarly as in our previous work [26].

Theorem 3.1 (Unknown targets temporal causal DAG identification). *Suppose $\mathcal{I}^*$ is such that $\mathcal{I}_1^* := \emptyset$. Let $\mathcal{G}^*$ be the ground truth temporal DAG and $(\hat{\mathcal{G}}, \hat{\mathcal{I}}) \in argmax_{\mathcal{G}\in DAG,\mathcal{I}}\mathcal{S}(\mathcal{G},\mathcal{I})$. Under the assumptions that: 1) the density model has enough capacity to represent the ground truth distributions; 2) $\mathcal{I}^*$-faithfulness holds; 3) the density model is strictly positive; 4) the ground truth densities $p^{(q)}$ have finite differential entropy. For $\lambda, \lambda_R > 0$ small enough, $\hat{\mathcal{G}}$ is $\mathcal{I}^* - Markov$ equivalent to $\mathcal{G}^*$ and $\hat{\mathcal{I}} = \mathcal{I}^*$.*

*3.1.4 Maximize the score.* Subsequently, to allow the gradient-based stochastic optimization process, we relax the above score by taking $M^{\mathcal{G}}$ and $R^{\mathcal{I}}$ as a random matrix respectively, where $M_{ij}^{\mathcal{G}} \sim B(1, \sigma(\alpha_{ij}))$ and $R_{qj}^{\mathcal{I}} \sim B(1, \sigma(\beta_{qj}))$, $B$ represents the Bernoulli distribution, $\sigma$ is the sigmoid function and $\alpha_{ij}, \beta_{qj}$ are scalar parameters. We group these $\alpha_{ij}$s into a matrix $\Lambda \in \mathbb{R}^{(p+1)d\times(p+1)d}$, and $\beta_{kj}$s into a matrix $\Gamma \in \mathbb{R}^{Q\times(p+1)d}$. After that, we rely on *augmented Lagrangian procedure* [61] to maximize the following score:

$$\hat{\mathcal{S}}(\Lambda, \Gamma) := \sup_{\phi} \mathop{\mathbb{E}}_{M\sim\sigma(\Lambda)} \left[ \mathop{\mathbb{E}}_{R\sim\sigma(\Gamma)} \left[ \sum_{q=1}^{Q} \mathop{\mathbb{E}}_{X\sim p^{(q)}} \log f^{(q)} \left( X; M, R^{\mathcal{I}^*}, \phi \right) - \lambda \|M\|_0 - \lambda_R \|R\|_0 \right] \right], \quad (5)$$

under the acyclicity constraint:

$$\sup_{\Lambda,\Gamma} \hat{\mathcal{S}}(\Lambda, \Gamma), \text{s.t.} \operatorname{Tr} e^{\sigma(\Lambda)} - (p+1)d = 0. \quad (6)$$

Moreover, as for gradient of the score w.r.t. $\alpha_{ij}$ and $\beta_{qj}$, following the general dealing method in continuous optimization for causal discovery [21, 32], we estimate $\Lambda$ and $\Gamma$ by *Straight-Through Gumbel estimator*, which means that Bernoulli samples are used in forward pass and Gumbel-Softmax samples are used in backward pass.

Overall, the learnable parameters in the process are $\phi$,$\Lambda$, $\Gamma$, and the estimated adjacency matrix reflecting temporal causal relations is $\sigma(\Lambda)$ and the estimated potential interventional family is $\sigma(\Gamma)$.

Since we only focus on influences on $X_1, \ldots, X_d$ from other variables, we set $\Lambda[:, d+1 : (p+1)d]$, i.e. the meaningless part $M^{\mathcal{G}}[:, d+1 : (p+1)d]$, to zero before training.

## 3.2 LLM-guided Meta Initialization

In this section, we introduce the detailed method of using LLM to bring in domain knowledge and extra prior information [9, 34] so that the potential temporal causal relations are obtained to guide the data-driven optimization process.

*Prompts of LLMs.* We describe the system into prompts as queries to the LLMs for possible temporal causal relationships. We implement strategies to mitigate biases, notably through the use of tailored prompts shown in Table 1, where the underlined units are indispensable and the others are optional. Prompt D is designed to clarify the causal discovery tasks for the LLMs and align them with the domain knowledge inherent to the LLMs themselves. Prompt I further introduces data and domain knowledge consistent with the subsequent causal discovery tasks to the LLMs, such as context descriptions, physical structures, and generating rules, to ensure that the meta-initialization process is as unbiased as possible.

**Table 1: Example prompt for LLM-guided Meta Initialization.**

**Prompt D**efinition

`Role`: "You are an exceptional temporal causal discovery analyzer, with in-depth domain knowledge in . . . (e.g. the intelligent operation and maintenance of data center air-conditioning systems)."
`Introduction`: "A directed temporal causal relationship between variables xu and xv can be represented as a tuple (xu, xv, t), signifying that the variable xu, lagging t time units, causally influences the current state of variable xv. The tuple (xu, xv, 0) denotes contemporaneous causality if t=0; if t>0, the tuple (xu, xv, t) indicates time-lagged causality. Note that when t=0, i.e. in (xu, xv, 0), xu and xv must be different variables, as intra-slice self-causality is not considered. Also, (xu, xv, 0) and (xu, xv, t) for t>0 have the possibility to coexist, suggesting that contemporaneous and time-lagged causality between two variables might simultaneously occur sometimes. Our task is to unearth all the possible temporal causal relationships among variables, grounded on the subsequent information."

**Prompt I**nformation

`Domain knowledge`: Depending on the application scenarios, it might be: 1) A *context description* of a specific industrial scenario or AIOps scenario. 2) The *physical structure* of the system, containing the *location* information of each entity or variable. 3) The abstract *generating rules* of time series.
`Data`: Providing a piece of past time series may help LLM understand the variables and their relations better.

**Prompt C**ausal Discovery in Temporal Domain

`Task`: "Please identify all temporal causal relations among the $n$ variables $(x_1, \ldots, x_n)$, considering only contemporaneous and $p$ time-lagged causality. Conclude your response with the full answer as a Python list of tuples (xu, xv, t) after 'Answer:'. Don't simplify and just give me some examples. You should cover all possible relationships in your answer."

**Prompt H**int

`Implication`: We can offer a thinking path to the LLM model as a guide of how we want it to utilize the prior information we gave to it or its own knowledge and deduct the answer.
`Chain of Thought (CoT)`: "Proceed methodically, step by step." By simply adding a zero-shot CoT prompt, the LLM model could output its answer with its path of thought, making the process more interpretable and easy for humans to understand, check, and correct immediately [29, 48]. Furthermore, if we can provide an example that contains the correct thought path and result as input, the one-shot CoT may achieve an even better result.

Using tuples provided by LLM, we construct the adjacency matrix $M_0^{\mathcal{G}}$ to denote the learned causal structure. These results represent potential causality based on domain information, not guaranteed causal definitions. *The subsequent score-based causal discovery optimization ensures the final results conform to causal definitions.* We incorporate meta-initialization information as follows:

*Guide temporal causal discovery from data.* By initiating $M^{\mathcal{G}}$ or to say $\Lambda$ of the module score-based temporal causal discovery as the results $M_0^{\mathcal{G}}$ from the module LLM-guided meta initialization before training and optimization, we not only lead the direction of the data-driven optimization process but also guarantee the theoretical integrity of the temporal causal discovery from interventional data.

*Extract weight.* Eventually, we obtain the estimated full weight matrix $\hat{F} = \sigma(\Lambda)$ of graph $\mathcal{G}$. Then, we form the adjacency matrix $\hat{M}^{\mathcal{G}}$ of the graph by adding an edge whenever $\sigma(\Lambda) > 0.5$ is acyclic. Finally, we can extract the intra-slice matrix $\hat{W} = \hat{M}^{\mathcal{G}}[1 : d, 1 : d]$ and inter-slice matrix $\hat{A}_k = \hat{M}^{\mathcal{G}}[kd+1 : (k+1)d, 1 : d]$ for each time lag $k = 1, \ldots, p$. They reflect causal relations of these $d$ variables in both a contemporary and time-lagged manner.

The overall framework and algorithm of **RealTCD** are summarized in Figure 1 and Algorithm 1 respectively.

**Algorithm 1** The overall algorithm for **RealTCD**

```
Require: All kinds of prior knowledge
Generate prompt as described in Table 1
Obtain causal results from LLM and transfer into matrix $M_0^{\mathcal{G}}$
Require: $M_0^{\mathcal{G}}$, hyperparameter $\lambda$, $\lambda_R$, hyperparameter for augmented Lagrangian process
Transfer the constrained problem defined by equation (5), (6) into the form of unconstrained problem following augmented Lagrangian
Initialize $\Lambda_0$ by $M_0^{\mathcal{G}}$, and also initialize $\phi_0, \Gamma_0$, max_iteration $U$, Lagrangian multiplier $\gamma_0$ and penalty coefficient $\mu_0$
while $0 \le t \le U$ and $h(\Lambda) := \operatorname{Tr} e^{\sigma(\Lambda)} - (p+1)d > 10^{-8}$ do
    Solve the $t$th unconstrained subproblem using stochastic gradient descent algorithm (we use RMSprop)
    Get sub solution $\phi_t^*, \Lambda_t^*, \Gamma_t^*$, and initialize $\phi_{t+1}, \Lambda_{t+1}, \Gamma_{t+1}$ by them
    Update $\gamma_{t+1}$ and $\mu_{t+1}$
    $t = t + 1$
end while
Form causal graph by adding an edge whenever $\sigma(\Lambda) > 0.5$ is acyclic
```

# 4 Experiments

## 4.1 Setups

*4.1.1 Baselines.* To evaluate the effectiveness of our method, we compare with the following models as baselines: **DYNOTEARS** [33], **PCMCI** [38], **TECDI** [26], **NeuralGC** [40]. Since NeuralGC only learns contemporaneous relationships, we use $W_{\text{full}}$ defined as below to compress both contemporaneous and time-lagged causal relationships learnt from RealTCD into a single metric for comparison with NeuralGC, which does not differentiate these 2 types.

$$W_{\text{full}}(i,j) = \begin{cases} 1, & W(i,j) + \sum_{k=1}^{p} A_k(i,j) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

This formulation captures the entire causal relationship from $i$ to $j$, assuming a relationship exists if any contemporaneous or time-lagged relationship is present.

The distinction in method selection across different settings was intentional and aligned with the capabilities of each algorithm: Since PCMCI and DYNOTEARS are designed for causal structure learning from observational data and can not incorporate interventional data, we included them in the "Unknown" interventional targets setting. As for TECDI, we use the same data with RealTCD. For baselines that are unable to bring in interventional data, we ensure a fair comparison by keeping the overall sample size consistent across all models, while using only observational data (or normal data in real datasets) for those baselines.

*4.1.2 Synthetic datasets.* We generate temporal data in two steps:

- Sample intra DAG and inter DAG following the *Erdős-Rényi* scheme, then sample parameters in weighted adjacency matrix, where elements in intra-slice matrix $W$ are uniformly from $[-1.0, -0.25] \cup [0.25, 1.0]$ and elements in inter-slice matrixes $A_k$ are uniformly from $[-1.0\alpha, -0.25\alpha] \cup [0.25\alpha, 1.0\alpha], \alpha = 1/\eta^k, \eta \ge 1, k = 1, \ldots, p$.
- Generate time series consistent with the sampled weighted graph following the standard structural vector autoregressive (SVAR) model[37]: $Y_0 = Y_0W + Y_1A_1 + \cdots + Y_pA_p + Z$, where $Z$ is random variables under the normal distribution. Then, sample interventional targets from nodes in $Y_0$, and generate perfect interventional data by cutting off the dependency of intervened nodes on their parents, i.e. setting $W_{ij}$ and $A_{kij}$ to zero, where $x_j$ is the variable in interventional targets and $x_i \in x_{\pi_j^{\mathcal{G}}}$.

Before training, all data are normalized by subtracting the mean and dividing by the standard deviation. We experimented on two simulated datasets: Dataset 1 contains 5 nodes, their 1 time-lagged variables and 5 different interventional targets. Dataset 2 contains 10 nodes, their 1 time-lagged variables and 10 different interventional targets. We initially chose to limit our datasets to a time delay of 1 to facilitate the evaluation and presentation. However, it is indeed feasible to increase the time delay of the data.

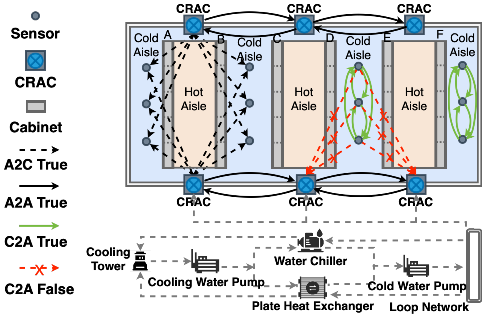


**Figure 2: A typical data center cooling system diagram.**

*4.1.3 Real-world data center application.* In contemporary data centers, IT equipment stability is crucial. Sophisticated air conditioning systems manage heat, maintaining a consistent temperature. Figure 2 shows a typical data center room organized with rows of equipment (A, B, C, D, etc.), separated by barriers to prevent hot and cold air mixing. Computer Room Air Conditioners (CRACs) on both sides of the room create a closed-loop system to maintain stable conditions. Multiple sensors in the cold aisle provide real-time temperature data, essential for ensuring continuous cold air delivery and stable IT operation.

When an anomaly occurs at a monitoring point, we can find the root cause of the anomaly based on the causal relationships among these entities we have learned through our **RealTCD** framework.

*Data acquisition.* The data used in this study was obtained from a specific data center at Alibaba. It covers monitoring data from a cooling system of a particular room from January 1st, 2023 to May 1st, 2023, and includes 38 variables in total. These variables comprise *18 cold aisle temperatures* from sensors and *20 air conditioning supply temperatures* from CRACs. We collected several time series during normal as well as abnormal states. For the latter, data was sampled within 20 minutes of the occurrence of the abnormality, with each sampling interval being 10 seconds. Anomaly points were identified by learning the normal distribution range from historical data, using the $n$-$\sigma$ method. Any data points that fall outside of the $n$-$\sigma$ range (e.g., 3 to 5) of itself are extracted as anomaly time points.

*4.1.4 Evaluation metrics. For synthetic datasets*, we leverage two metrics to evaluate the performance of learning causal graph: i) structural Hamming distance (**SHD**), which calculates the number of different edges (either reversed, missing or redundant) between two DAGs; ii) structural interventional distance (**SID**), which represents the difference between two DAGs according to their causal inference conditions[36]. *For real-world datasets*, given the absence

**Table 2: Results on synthetic and real-world datasets.**

| | | | Synthetic Dataset 1 | | Synthetic Dataset 2 | | Real-world Dataset | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Targets | Causality | Method | SHD ↓ | SID ↓ | SHD ↓ | SID ↓ | All Edges | C2A False ↓ | A2C True ↑ | A2A True ↑ | C2C True ↑ |
| Known | intra+inter | TECDI | 1.6 ± 2.3 | 1.8 ± 2.4 | 3.9 ± 5.1 | 10.6 ± 10.4 | 85.6 ± 12.5 | 4.0 ± 6.9 | 5.8 ± 3.2 | 0.7 ± 1.3 | 2.8 ± 1.3 |
| | | RealTCD | **1.2** ± **2.8** | **1.6** ± **3.1** | **2.2** ± **4.3** | **9.0** ± **10.5** | 79.5 ± 8.0 | **0.0** ± **0.0** | **6.1** ± **1.3** | **10.5** ± **2.2** | **5.7** ± **3.0** |
| Unknown | intra+inter | DYNOTEARS | 20.4 ± 2.4 | 38.6 ± 3.7 | 36.0 ± 5.2 | 118.6 ± 20.7 | 38.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 |
| | | PCMCI | 18.1 ± 4.4 | 24.1 ± 3.1 | 62.0 ± 17.2 | 118.3 ± 24.1 | 181.6 ± 32.3 | 32.9 ± 5.0 | **13.2** ± **2.9** | 7.5 ± 5.9 | 5.0 ± 2.5 |
| | | TECDI | 11.9 ± 4.8 | 17.7 ± 8.5 | 27.4 ± 10.7 | 60.8 ± 25.7 | 57.6 ± 10.7 | 2.4 ± 1.5 | 1.5 ± 1.8 | 1.8 ± 1.5 | 0.4 ± 1.0 |
| | | RealTCD | **9.9** ± **2.9** | **16.1** ± **6.2** | **7.1** ± **4.4** | **18.5** ± **11.0** | 51.7 ± 8.0 | **0.0** ± **0.0** | 1.5 ± 1.0 | **14.0** ± **1.6** | **8.4** ± **1.9** |
| | intra | NeuralGC | 14.9 ± 2.3 | 20.0 ± 0.0 | 31.3 ± 4.7 | 85.5 ± 6.4 | 104.1 ± 41.1 | 25.7 ± 19.5 | **1.9** ± **4.0** | 5.2 ± 3.7 | 4.0 ± 3.7 |
| | | RealTCD | **6.3** ± **1.8** | **18.1** ± **3.8** | **5.6** ± **3.8** | **75.3** ± **11.9** | 38.6 ± 6.0 | **0.0** ± **0.0** | 1.3 ± 0.7 | **12.0** ± **1.0** | **6.8** ± **1.0** |

of ground truth DAGs in real case, we employ four performance metrics based on expert knowledge to assess algorithms' efficacy. These metrics mainly consider the physical location relationships within the room, shown in figure 2. i) **A2C True** counts the correctly identified causal edges from *air conditioning supply temperatures* ($A$) to the *temperatures of* adjacent *cold aisles* ($C$), assuming that $A$ influences $C$. ii) **A2A True** counts the correctly identified causal edges between adjacent $A$ units, assuming mutual influences among them. iii) **C2C True** counts the correctly identified causal edges between $C$ in the same column, assuming direct interactions among them. iv) **C2A False** counts the incorrectly identified causal edges from $C$ to $A$, as such causality is implausible given that downstream variables $C$ cannot influence upstream variables $A$.[1]

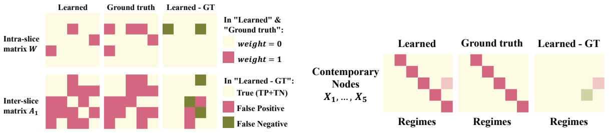
## 4.2 Main Results

Results on synthetic and real-world datasets are reported in Table 2.

*4.2.1 On synthetic datasets.* On both datasets, our method outperforms baseline models on SHD and SID metrics, with small standard deviations. For the dataset with 10 nodes, the improvement is more pronounced, highlighting our method's advantages in handling a **large number of variables**, making optimization more **focused** and **effective**. RealTCD consistently achieves better performance in each setting. "Known" targets setting performs better than the "Unknown" targets setting since the former employs ground-truth interventional target labels for training. However, since ground-truth targets are often unavailable in real scenarios, RealTCD's effectiveness with unknown targets is highly practical.

Figure 3a and Figure 3b display example results of the temporal causal DAG and interventional targets obtained through RealTCD. In Figure 3b, each row represents a contemporary node, and each column represents a regime corresponding to each interventional family: the first for observational data and the next five for different interventional data. Pink cells indicate the learned or ground truth interventional targets in each regime. These figures demonstrate our method's ability to accurately identify both the temporal causal graph and interventional targets.

*4.2.2 On real-world data center datasets.* Our method achieves fewer C2A False and learns more A2A True and C2C True relations. Introducing prior domain knowledge with the LLM module effectively understands the upstream and downstream relationships in the system architecture, avoiding downstream influence on upstream variables and keeping C2A False at zero in the subsequent optimization. Additionally, the standard deviation of each metric is small. Therefore, RealTCD outperforms other approaches in industrial scenarios by utilizing rich information from interventional data and prior domain knowledge from textual information.



(a) DAG results. (b) Interventional targets.

**Figure 3: Showcases of the results on synthetic data.**

RealTCD in the "Unknown" targets setting, especially designed in our paper, achieves the best performance across all settings, even outperforming the "Known" targets setting. This is because the interventional targets provided by users are often not ground truth (people can only detect anomalous variables, but cannot confirm whether they are the ground truth interventional targets), which could potentially mislead the learning of algorithm. This further underscores the importance of using our method based on interventional data with unknown targets to deal with real-world cases.

## 4.3 Deeper Analysis

*4.3.1 Ablation studies.* We conduct ablation studies to verify the effectiveness of the proposed 2 modules in **RealTCD** as in Table 3.

'Variant1' is 'RealTCD w/o intervention', which **removes the interventional module** and uses only observational data of the same sample size. It tends to output more edges, making comparisons unfair, so we calculate the ratio of labeled edges to total edges learned for further evaluation. Removing the interventional module significantly decreases performance, especially for the A2A True and C2C True ratios, confirming the effectiveness of our score-based temporal causal discovery from interventional data.

'Variant2' is 'RealTCD w/o LLM', which **removes the LLM-guided meta initialization**. Performance drops significantly without this module, even reaching zero for the A2A True metric. This shows that the LLM-guided meta initialization effectively utilizes

[1] Across all tables, ↑ denotes the higher the better, and ↓ denotes the lower the better. The best results of a particular setting are in bold.

prior information from system textual data, greatly enhancing temporal causal discovery in real-world scenarios. Integrating this module is crucial for providing a well-informed starting point, reducing biases, and improving causal inference precision.

**Table 3: Ablation studies of modules on real-world dataset.**

| Method | All Edges | C2A False↓ | A2C True↑ | A2A True↑ | C2C True↑ |
|---|---|---|---|---|---|
| RealTCD | 51.7 ± 8.0 | 0.0 ± 0.0 | 1.5 ± 1.0 | 14.0 ± 1.6 | 8.4 ± 1.9 |
|  | % | 0.0 ± 0.0 | 3.1 ± 2.4 | **27.3** ± **2.7** | **16.5** ± **3.6** |
| Variant1 | 304.3±18.7 | 0.0 ± 0.0 | 43.3 ± 7.3 | 16.5 ± 1.6 | 14.8 ± 1.1 |
|  | % | 0.0 ± 0.0 | **14.2** ± **1.8** | 5.4 ± 0.5 | 4.9 ± 0.4 |
| Variant2 | 38.8 ± 1.9 | 0.0 ± 0.0 | 0.1 ± 0.3 | 0.0 ± 0.0 | 0.4 ± 0.7 |
|  | % | 0.0 ± 0.0 | 0.2 ± 0.7 | 0.0 ± 0.0 | 1.0 ± 1.6 |

*4.3.2 Different LLMs and prompts.* We tested different LLMs and prompts on real-world datasets [19, 62]. GPT-4, especially with prompts containing implication and CoT, achieved better and more stable results. When human ideas are unclear, zero-shot CoT can effectively leverage LLMs' domain knowledge, providing solutions that reveal the thinking process and enhance interpretability.

**Table 4: Comparisons of different LLMs and prompts.**

| Models | GPT-4 |  |  | GPT-3.5-turbo-instructor |  |  |
|---|---|---|---|---|---|---|
| Prompt No. | 0 | 1 | 2 | 0 | 1 | 2 |
| `Implication` | ✓ |  |  | ✓ |  |  |
| `CoT` | ✓ | ✓ |  | ✓ | ✓ |  |
| C2A False ↓ | 0 | 0 | 0 | 0 | 0 | 0 |
| A2C True ↑ | 60 | **68** | 18 | 8 | 8 | 32 |
| A2A True ↑ | **36** | 0 | 0 | 0 | 18 | 0 |
| C2C True ↑ | **36** | 0 | 12 | 0 | 0 | 0 |

# 5 Discussion

## 5.1 Motivation and Strengths of Using LLMs

In industrial systems, operations come with extensive textual documentation like logs and manuals, providing valuable insights into system behavior. LLMs excel at mining this textual data to inform causal analysis, directly connecting text and system operations. We outline **the irreplaceable strengths of using LLMs** in RealTCD over traditional deep learning approaches as follows: *Handling Textual Information*: Traditional methods often ignore the rich textual data in real-world systems, while LLMs process and utilize it to enhance causal discovery accuracy. Though conventional language models can also process text, they lack the advanced capabilities of LLMs in in-context learning [8], which provides LLMs with superior generalization and flexibility, as well as powerful zero-shot and few-shot learning abilities [7, 25]. *Integration of Domain Knowledge*: LLMs assimilate user inputs (e.g. the structure of a particular system) and integrate extensive domain-specific knowledge embedded within the LLMs (e.g. the operation law of a system), covering areas unfamiliar to users. This is crucial for accurate causal inference in complex systems. *LLM-guided Meta-Initialization*: This module significantly improves causal discovery quality by using meta-knowledge from LLMs to narrow the scope initially, unlike traditional methods that start with broad assumptions, which can lead to suboptimal local solutions and high variance in results. LLMs' suitability for enhancing causal discovery is supported by literature demonstrating their effectiveness in complex inference tasks [4, 11, 20, 24, 43].

## 5.2 Practical Implications

In our paper, we highlight the application of temporal causal discovery within AIOps to enhance systems' monitoring, troubleshooting, and predictive capabilities, crucial for tasks such as anomaly detection [54], root cause analysis [45], failure prediction, and system optimization. It is also helpful in various fields such as finance [22], healthcare [15, 31, 42, 56], and social sciences [16] by uncovering the dynamic interrelations between variables over time.

# 6 Related Work

*Causal Discovery in Temporal Domain.* Interventional data greatly aids in identifying causal structures [13, 14, 41, 60], but designing experiments and collecting data is challenging. Jaber et al. [18] used a Ψ-Markov property for learning causal graphs with latent variables. [1] proposed a $p$-collider-based algorithm to recover causal graphs with minimal intervention costs. [6] used interventional data and neural architectures for causality detection. [26] focused on temporal causality from observational data but required intervention labels, limiting practicality. We focus on discovering temporal causal relationships without intervention labels.

*LLMs for Causal Discovery.* Recent works explored LLMs for causal discovery[3, 5, 12, 17, 20, 23, 27, 39, 44, 49] and dynamic graphs[46, 47, 50–52, 55, 58]. Ban et al. [4] showed LLMs' potential in causal inference. Chan et al. [10] assessed ChatGPT's ability to capture temporal and causal relations. Chen et al. [11] used LLM-driven knowledge to reduce biases in causal learning. [24] highlighted LLMs' roles in causal discovery and reasoning. Long et al. [28] investigated LLMs' understanding of causal relationships in medical contexts. We are the first to use LLMs for temporal causal discovery to the best of our knowledge.

# 7 Conclusion

In this paper, we propose the **RealTCD** framework, a novel approach for temporal causal discovery in AIOps that bypasses the need for interventional targets and leverages system textual information. Our method, featuring score-based temporal causal discovery and LLM-guided meta-initialization, outperforms existing baselines on both simulated and real-world datasets. The results underscore **RealTCD**'s potential to enhance causal analysis in complex IT systems and suggest further research in AIOps.

## Acknowledgments

This work was supported in part by the National Key Research and Development Program of China No. 2020AAA0106300, National Natural Science Foundation of China (No. 62222209, 62250008, 62102222), Beijing National Research Center for Information Science and Technology Grant No. BNR2023RC01003, BNR2023TD03006, Alibaba Innovative Research Program and Beijing Key Lab of Networked Multimedia.

## References

[1] Raghavendra Addanki, Shiva Kasiviswanathan, Andrew McGregor, and Cameron Musco. 2020. Efficient intervention design for causal discovery with latents. In *International Conference on Machine Learning*. PMLR, 63–73.

[2] Vijay Arya, Karthikeyan Shanmugam, Pooja Aggarwal, Qing Wang, Prateeti Mohapatra, and Seema Nagar. 2021. Evaluation of causal inference techniques for AIOps. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*. 188–192.

[3] 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* (2023).

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

[5] Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2023. Towards LLM-guided Causal Explainability for Black-box Text Classifiers. https://api.semanticscholar.org/CorpusID:262459118

[6] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. 2020. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems* 33 (2020), 21865–21877.

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

[8] Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs Are Few-Shot In-Context Low-Resource Language Learners. *arXiv preprint arXiv:2403.16512* (2024).

[9] Alessandro Castelnovo, Riccardo Crupi, Fabio Mercorio, Mario Mezzanzanica, Daniele Potertì, and Daniele Regoli. 2024. Marrying LLMs with Domain Expert Validation for Causal Graph Generation. (2024).

[10] Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827* (2023).

[11] Lyuzhou Chen, Taiyu Ban, Xiangyu Wang, Derui Lyu, and Huanhuan Chen. 2023. Mitigating Prior Errors in Causal Structure Learning: Towards LLM driven Prior Knowledge. *arXiv preprint arXiv:2306.07032* (2023).

[12] Kai-Hendrik Cohrs, Emiliano Diaz, Vasileios Sitokonstantinou, Gherardo Varando, and Gustau Camps-Valls. 2023. Large Language Models for Constrained-Based Causal Discovery. In *AAAI 2024 Workshop on"Are Large Language Models Simply Causal Parrots?"*.

[13] Frederick Eberhardt. 2012. Almost Optimal Intervention Sets for Causal Discovery. arXiv:1206.3250 [cs.AI]

[14] Tian Gao, Debarun Bhattacharjya, Elliot Nelson, Miao Liu, and Yue Yu. 2022. IDYNO: Learning Nonparametric DAGs from Interventional Dynamic Data. In *International Conference on Machine Learning*. PMLR, 6988–7001.

[15] Chang Gong, Di Yao, Chuzhe Zhang, Wenbin Li, Jingping Bi, Lun Du, and Jin Wang. 2023. Causal discovery from temporal data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5803–5804.

[16] Anna Grzymala-Busse. 2011. Time will tell? Temporality and the analysis of causal mechanisms and processes. *Comparative Political Studies* 44, 9 (2011), 1267–1297.

[17] George Gui and Olivier Toubia. 2023. The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective. *ArXiv* abs/2312.15524 (2023). https://api.semanticscholar.org/CorpusID:266133974

[18] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. 2020. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems* 33 (2020), 9551–9561.

[19] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2024. Cladder: A benchmark to assess causal reasoning capabilities of language models. *Advances in Neural Information Processing Systems* 36 (2024).

[20] 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* (2024).

[21] Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. 2022. Structural agnostic modeling: Adversarial learning of causal graphs. *The Journal of Machine Learning Research* 23, 1 (2022), 9831–9892.

[22] Prabhanjan Kambadur, Aurélie C Lozano, and Ronny Luss. 2016. Temporal causal modeling. *Financial Signal Processing and Machine Learning* (2016), 41–66.

[23] Tejas Kasetty, Divyat Mahajan, Gintare Karolina Dziugaite, Alexandre Drouin, and Dhanya Sridhar. 2024. Evaluating Interventional Reasoning Capabilities of Large Language Models. https://api.semanticscholar.org/CorpusID:269004745

[24] 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* (2023).

[25] Guozheng Li, Peng Wang, Jiajun Liu, Yikai Guo, Ke Ji, Ziyu Shang, and Zijie Xu. 2024. Meta In-Context Learning Makes Large Language Models Better Zero and Few-Shot Relation Extractors. *arXiv preprint arXiv:2404.17807* (2024).

[26] Peiwen Li, Yuan Meng, Xin Wang, Fang Shen, Yue Li, Jialong Wang, and Wenwu Zhu. 2023. Causal Discovery in Temporal Domain from Interventional Data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4074–4078.

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

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

[29] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379* (2023).

[30] Søren Wengel Mogensen, Karin Rathsman, and Per Nilsson. 2023. Causal discovery in a complex industrial system: A time series benchmark. *arXiv preprint arXiv:2310.18654* (2023).

[31] Narmada Naik, Ayush Khandelwal, Mohit Joshi, Madhusudan Atre, Hollis Wright, Kavya Kannan, Scott Hill, Giridhar Mamidipudi, Ganapati Srinivasa, Carlo Bifulco, et al. 2023. Applying Large Language Models for Causal Structure Learning in Non Small Cell Lung Cancer. *arXiv preprint arXiv:2311.07191* (2023).

[32] Ignavier Ng, Shengyu Zhu, Zhuangyan Fang, Haoyang Li, Zhitang Chen, and Jun Wang. 2022. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. SIAM, 424–432.

[33] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. 2020. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1595–1605.

[34] Nick Pawlowski, James Vaughan, Joel Jennings, and Cheng Zhang. 2023. Answering causal questions with augmented llms. (2023).

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

[36] Jonas Peters and Peter Bühlmann. 2015. Structural intervention distance for evaluating causal graphs. *Neural computation* 27, 3 (2015), 771–799.

[37] Juan F Rubio-Ramirez, Daniel F Waggoner, and Tao Zha. 2010. Structural vector autoregressions: Theory of identification and algorithms for inference. *The Review of Economic Studies* 77, 2 (2010), 665–696.

[38] Jakob Runge. 2020. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 1388–1397.

[39] 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* (2024).

[40] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. 2021. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2021), 4267–4279.

[41] Jin Tian and Judea Pearl. 2013. Causal discovery from changes. *arXiv preprint arXiv:1301.2312* (2013).

[42] Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. 2024. Automating Psychological Hypothesis Generation with AI: Large Language Models Meet Causal Graph. *arXiv preprint arXiv:2402.14424* (2024).

[43] 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* (2023).

[44] Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. 2023. Causal Inference Using LLM-Guided Discovery. *ArXiv* abs/2310.15117 (2023). https://api.semanticscholar.org/CorpusID:264591509

[45] Dongjie Wang, Zhengzhang Chen, Yanjie Fu, Yanchi Liu, and Haifeng Chen. 2023. Incremental Causal Graph Learning for Online Root Cause Analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2269–2278.

[46] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. 2017. Community preserving network embedding. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[47] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The world wide web conference*. 2022–2032.

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[49] Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. 2022. Probing for correlations of causal facts: Large language models and causality.

(2022).

[50] Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853* (2024).

[51] Siheng Xiong, Yuan Yang, Faramarz Fekri, and James Clayton Kerce. 2024. TILP: Differentiable learning of temporal logical rules on knowledge graphs. *arXiv preprint arXiv:2402.12309* (2024).

[52] Siheng Xiong, Yuan Yang, Ali Payani, James C Kerce, and Faramarz Fekri. 2024. Teilp: Time prediction over knowledge graphs via logical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 16112–16119.

[53] Wenzhuo Yang, Kun Zhang, and Steven Hoi. 2022. A Causal Approach to Detecting Multivariate Time-series Anomalies and Root Causes. (2022).

[54] Wenzhuo Yang, Kun Zhang, and Steven CH Hoi. 2022. Causality-based multivariate time series anomaly detection. *arXiv preprint arXiv:2206.15033* (2022).

[55] Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. 2023. Understanding Causality with Large Language Models: Feasibility and Opportunities. arXiv:2304.05524 [cs.LG]

[56] Zeyang Zhang, Xingwang Li, Fei Teng, Ning Lin, Xueling Zhu, Xin Wang, and Wenwu Zhu. 2023. Out-of-Distribution Generalized Dynamic Graph Neural Network for Human Albumin Prediction. In *IEEE International Conference on Medical Artificial Intelligence*.

[57] Ziwei Zhang, Xin Wang, Zeyang Zhang, Peng Cui, and Wenwu Zhu. 2021. Revisiting Transformation Invariant Geometric Deep Learning: Are Initial Representations All You Need? *arXiv preprint arXiv:2112.12345* (2021).

[58] Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Yijian Qin, Simin Wu, and Wenwu Zhu. 2023. LLM4DyG: Can Large Language Models Solve Spatio-Temporal Problems on Dynamic Graphs? *arXiv preprint arXiv:2310.17110* (2023).

[59] Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, and Wenwu Zhu. 2023. Out-of-Distribution Generalized Dynamic Graph Neural Network with Disentangled Intervention and Invariance Promotion. *arXiv preprint arXiv:2311.14255* (2023).

[60] Zeyang Zhang, Xin Wang, Ziwei Zhang, Zhou Qin, Weigao Wen, Hui Xue, Haoyang Li, and Wenwu Zhu. 2023. Spectral Invariant Learning for Dynamic Graphs under Distribution Shifts. In *Advances in Neural Information Processing Systems*.

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

[62] Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. 2024. CausalBench: A Comprehensive Benchmark for Causal Learning Capability of Large Language Models. *arXiv preprint arXiv:2404.06349* (2024).