Causal Discovery through Synergizing Large Language Model and Data-Driven Reasoning

Huaming Du
School of Computing and Artificial Intelligence
Southwestern University of Finance and Economics
Chengdu, Sichuan, China

Yujia Zheng
Department of Philosophy
Carnegie Mellon University
Pittsburgh, Pennsylvania, United States

Baoyu Jing
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
Champaign, Illinois, United States

Yu Zhao
Financial Intelligence and Financial Engineering Key Laboratory
Southwestern University of Finance and Economics
Chengdu, Sichuan, China

Gang Kou
Xiangjiang Laboratory
Changsha, Hunan, China
Hunan University of Technology and Business
Changsha, Hunan, China
Southwestern University of Finance and Economics
Chengdu, Sichuan, China

Guisong Liu
Engineering Research Center of Intelligent Finance, Ministry of Education
Southwestern University of Finance and Economics
Chengdu, Sichuan, China

Tao Gu
Sichuan University
Chengdu, Sichuan, China
School of Business Administration
Southwestern University of Finance and Economics
Chengdu, Sichuan, China

Weimin Li
Department of Pulmonary and Critical Care Medicine
Frontiers Science Center for Disease-related Molecular Network
West China Hospital
Sichuan University
Chengdu, Sichuan, China

Carl Yang*
Department of Computer Science
Emory University
Atlanta, Georgia, United States

ABSTRACT

Revealing the underlying causal mechanisms in the real world is critical for scientific and technical progress. Despite advancements over the past decades, the lack of high-quality data and the inability of traditional causal discovery algorithms (TCDA) to fully comprehend the exact semantics of variables have long been major obstacles to the broader application of causal discovery. To address this issue, this paper proposes a novel causal modeling framework, **LLM-CD**, which integrates the metadata-based reasoning capabilities of large language models (LLMs) with the data-driven modeling abilities of TCDA for causal discovery. LLM-CD deeply couples the reasoning abilities of LLMs at various stages of TCDA, and enhances causal discovery through an iterative process. Due to the issues of overconfidence and hallucination in LLMs, LLM-CD quantifies and analyzes its uncertainty by incorporating evidence-based deep learning theory with the assumptions of TCDA. We utilize a large-scale de-identified real patient dataset provided by a hospital, a new dataset extracted from MIMIC-IV about the same disease (lung cancer), and two benchmark datasets to comprehensively evaluate LLM-CD. Extensive experimental results confirm the effectiveness and reliability of LLM-CD, with the highest improvement of 403.93% in the Recall and 25.77% in the Ratio metric across four datasets.

CCS CONCEPTS

• **Mathematics of computing → Causal networks**.

KEYWORDS

Causal discovery, Large language model, Healthcare

ACM Reference Format:
Huaming Du, Yujia Zheng, Baoyu Jing, Yu Zhao, Gang Kou, Guisong Liu, Tao Gu, Weimin Li, and Carl Yang. 2025. Causal Discovery through Synergizing Large Language Model and Data-Driven Reasoning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3711896.3736874

*corresponding author (j.carlyang@emory.edu)

KDD '25, August 3–7, 2025, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08…$15.00
https://doi.org/10.1145/3711896.3736874

1 INTRODUCTION

Causal discovery is crucial in both society and academia, providing valuable insights for policy-making, scientific research, and technological advancements [16, 27]. For example, in the healthcare field, causal discovery can identify potential causal relationships between a particular medication choice and disease treatment outcomes. The gold standard is to use randomized experiments, but this is often too costly or even unethical. Therefore, most Traditional Causal Discovery Algorithms (TCDA) primarily rely on analyzing observational data to reveal causal relationships. These methods mainly include constraint-based algorithm (e.g., PC [44] and CD-NOD [15]), score-based methods (e.g., GES [7]), and some Functional Causal Model (FCM)-based methods [33, 39]. However, in real-world complex systems, people often fail to collect and measure large quantities of high-quality, task-relevant data, and TCDA struggles to handle high-dimensional data and cannot fully understand the specific semantics of variables. For example, as shown in Figure 1(a), TCDA incorrectly concludes a causal relationship between lung cancer and organization name based solely on observational data. Furthermore, in TCDA, both constraint- and score-based methods ultimately results in a Markov equivalence class rather than a unique Directed Acyclic Graph (DAG), introducing uncertainty. For instance, a causal graph satisfying the condition $v_1 \perp v_3 \mid v_2$ can have three different structures: $v_1 \to v_2 \to v_3$, $v_1 \leftarrow v_2 \to v_3$, and $v_1 \leftarrow v_2 \leftarrow v_3$.

Causal discovery with LLMs has received widespread attention from the community. A series of early attempts suggest that LLMs can effectively understand the specific semantics of variables and leverage the learned knowledge to answer commonsense causal questions [1, 58]. For example, some works use LLMs to predict causal relationships [31, 52], employ LLMs as priors for data-driven causal discovery methods [3, 26], and assist downstream causal inference tasks by predicting the causal order of variables [47]. However, existing methods mainly use LLMs as a simple reasoner to directly infer causal relationships for given causal variables, often lacking an essential understanding of complex causal relationships. Moreover, these methods tend to focus more on the performance of LLMs in causal tasks, often overlooking the uncertainty issues arising from the models' overconfidence and hallucination [59]. For instance, as illustrated in Figure 1(a), directly asking LLMs about the causal relationship between potassium and lung cancer often leads to the incorrect conclusion that potassium is the cause of lung cancer.

Recent studies have shown that LLMs possess certain causal reasoning capabilities, enabling them to identify variable semantics and assist in causal discovery by integrating prior knowledge [1, 18, 24]. However, existing research is subject to the following limitations: *First*, most prior works [8, 20, 24, 31] are heuristic and lack cansal principles, directly using LLMs to dominate the entire causal discovery process. *Second*, most prior works [8, 31] rely on single-step queries to infer causal relationships between variables, without utilizing a multi-step iterative process. *Third*, Most studies [1, 8, 20, 31] overlook the issues of overconfidence and hallucination in LLMs and do not quantify the model's uncertainty.

To address these limitations, we propose a novel LLMs-based Causal Discovery Framework (LLM-CD). LLM-CD combines the

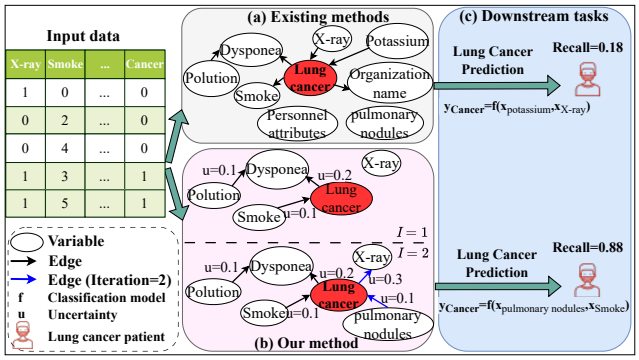


Figure 1: A toy example of causal discovery and downstream lung cancer prediction using tabular data. The main differences between LLM-CD and existing methods are: 1) a deep integration of LLM with TCDA methods, fully leveraging the strengths of both LLM and TCDA; 2) the design of an iterative mechanism to fully exploit the reasoning capabilities of LLM; 3) the adoption of evidence deep learning theory to thoroughly explore the uncertainty of the model.

data-driven modeling of the PC algorithm with the world-knowledge reasoning used by LLMs in causal discovery tasks, aiming to facilitate causal discovery and enhance the performance of downstream tasks. Specifically, we choose the PC algorithm [44] as the TCDA, because (1) it is generalizable to various causal discovery scenarios; (2) it is well supported by the causal principles; (3) it is a multi-stage method, allowing for integration with LLM in a divide-and-conquer manner. As for the integration between PC and LLM, we first use the LLMs to pre-screen variables based on downstream tasks, reducing redundant variables to improve algorithm efficiency. Next, we use the LLMs to assist the PC algorithm in conditional independence testing and edge orientation. Then, we apply Breadth-First Search (BFS) to detect cyclic structures in the obtained causal graph and use the LLMs to perform cycle elimination, ultimately resulting in a DAG. Based on this DAG, we perform downstream prediction tasks. We also design an iterative process to re-execute the above steps on misclassified samples to iteratively discover more accurate causal graph.

Overconfidence and hallucinations in LLMs [55], as well as the PC outputting Markov equivalence classes, limit their deployment in critical applications, particularly in fields such as healthcare, finance, and security. This makes *uncertainty estimation*[1] a crucial component in preventing potentially disastrous decisions based on outputs from existing methods. Therefore, by incorporating evidence deep learning theory [40], we analyze the uncertainty of our method under both white-box and black-box LLMs.

Extensive experiments on our own large-scale real-world hospital dataset WCHSU, a unique lung cancer dataset extracted from MIMIC-IV [21], and two benchmark datasets are conducted, aiming at DAG recovery and downstream prediction tasks. The comprehensive evaluation demonstrates the superiority of LLM-CD in comparison to SOTA methods. The results and discussions, along with ablation studies, human evaluations, LLM behavior experiments,

[1] uncertainty estimation refers to the inconsistency of the model's output results under different experimental setups. For example, in Figure 1(b), the causal relationship between X-ray and lung cancer varies with different prompts and data subsets.

uncertainty estimation, cost analysis, hyperparameter studies, and case studies, will be presented and analyzed in Section 4.

In summary, the main contributions are stated as follows:

• A unified causal discovery framework called LLM-CD, which integrates LLMs with TCDA to obtain more accurate causal graphs and empower downstream tasks, is proposed.

• At various stages of the PC algorithm, we combine the prior knowledge of LLMs with LLMs acting as priors and critics for causal discovery. Through an iterative process, we further unleash the causal reasoning capabilities of LLMs. Additionally, we incorporate evidence deep learning theory into LLM-CD to quantify the model's uncertainty.

• Extensive experimental results on various real-world medical and generic benchmark datasets demonstrate the superiority of LLM-CD.

2 RELATED WORK

In this section, we first introduce different causal discovery methods, followed by an overview of LLMs reasoning. Finally, we review the cutting-edge work on using LLMs to assist causal discovery.

2.1 Causal Discovery

Causal discovery aims to reveal unknown causal relations from observational data [19, 60], which is critical for both practical applications and scientific discoveries [38]. Current causal discovery strategies can be broadly classified into constraint-based methods, score-based methods, and functional causal model-based methods. Constraint-based methods have been employed for revealing information about the underlying causal structure through conditional independence relations in the data. Algorithms such as the PC algorithm and Fast Causal Inference (FCI) algorithm [45] are commonly used. Score-based algorithms like GES [7] aim to find the causal structure by optimizing a score function, such as the Bayesian Information Criterion. Functional causal model-based methods rely on specific parametric assumptions to infer directions from the independent noise condition. These methods provide asymptotically correct results, accommodating various data distributions and functional relationships.

Despite recent progress in both theoretical and empirical aspects [12], most TCDA still fail to accurately understand the specific semantics of variables and rely on large amounts of high-quality data. This reliance becomes particularly problematic when data quality is low and quantity is limited, as the corresponding identifiability guarantees are often restricted to the asymptotic case. This highlights a potential gap between theoretical assumptions and real-world applicability.

2.2 Reasoning with LLMs

LLMs have demonstrated significant performance improvements in various reasoning tasks through zero-shot or few-shot demonstrations [57]. LLMs are capable of acquiring and understanding commonsense knowledge about the world [5], and providing appropriate context as input can further unlock their potential [50]. However, studies have shown that LLMs may generate non-factual results [59], tend to learn shortcuts or dataset biases [28], and perform poorly in complex planning and reasoning tasks [5].

These limitations make it risky to rely solely on LLM reasoning results to derive rigorous conclusions. Therefore, instead of directly deriving results from LLMs, we deeply integrate LLMs with TCDA to achieve mutual enhancement. This approach enables the construction of more accurate causal graph structures, which in turn supports downstream prediction tasks.

2.3 Integration of LLM in Causal Discovery

Causal learning with LLMs has received much attention from the community [10, 23, 29, 49, 53, 54]. Kiciman et al. [23] find that LLMs can recover the pairwise causal relations relatively well. Some research [1, 31] propose to incorporate the causal discovery results by LLMs as a prior or constraint to improve the performance of data-driven causal discovery algorithms. However, Willig et al. [51] find that LLMs can not understand causality but simply retell the causal knowledge contained in the training data. Abdulaal et al. [1] proposed a causal modeling agent framework based on LLMs for causal discovery tasks, yet it requires additional deep chain graph models [9] for post-processing. Long et al. [32] believe that LLMs provide an exciting opportunity to supplement and accelerate DAG creation and are capable of building causal graphs with 3-4 nodes. However, Tu et al. [46] and Jin et al. [17] find that the performance of LLMs in more complex causal discovery remains limited as LLMs can hardly understand new concepts and knowledge.

The aforementioned debate highlights the limitations of directly using LLMs for causal discovery, which motivates us to deeply integrate existing causal discovery algorithms, rather than solely relying on LLMs, to learn causal relations.

3 METHODOLOGY

We introduce the LLM-CD framework (see Figure 2), which unifies modeling paradigms based on LLMs and data to infer the causal relationships between variables in the dataset, thereby supporting downstream tasks. The framework consists of six key steps: 1) initial variable screening, 2) skeleton construction, 3) edge orientation, 4) cycle removal, 5) iteration, and 6) uncertainty analysis. Please note that LLMs are regarded as large-scale background knowledge providers in this paper, and the core identifiability guarantee still relies on the classical PC framework.

3.1 Problem Definition

Causal discovery aims to infer causal relations among variables of interest $\mathcal{V}$ from the observational dataset, with the goal of constructing a causal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Based on $\mathcal{G}$, we focus on using the parent nodes of the target variable in the graph to predict the target variable, such as the lung cancer variable in Aisa dataset.

3.2 The Proposed Framework

1. Initial variable screening: In the process of causal discovery, existing causal discovery algorithms often have high complexity when dealing with large causal graphs. In addition, noise often exists in observational data, so we need to perform an initial screening of variables. By leveraging the prior knowledge of LLMs and providing appropriate prompts, LLMs can filter the input variables that are relevant to the target variable, thereby eliminating irrelevant

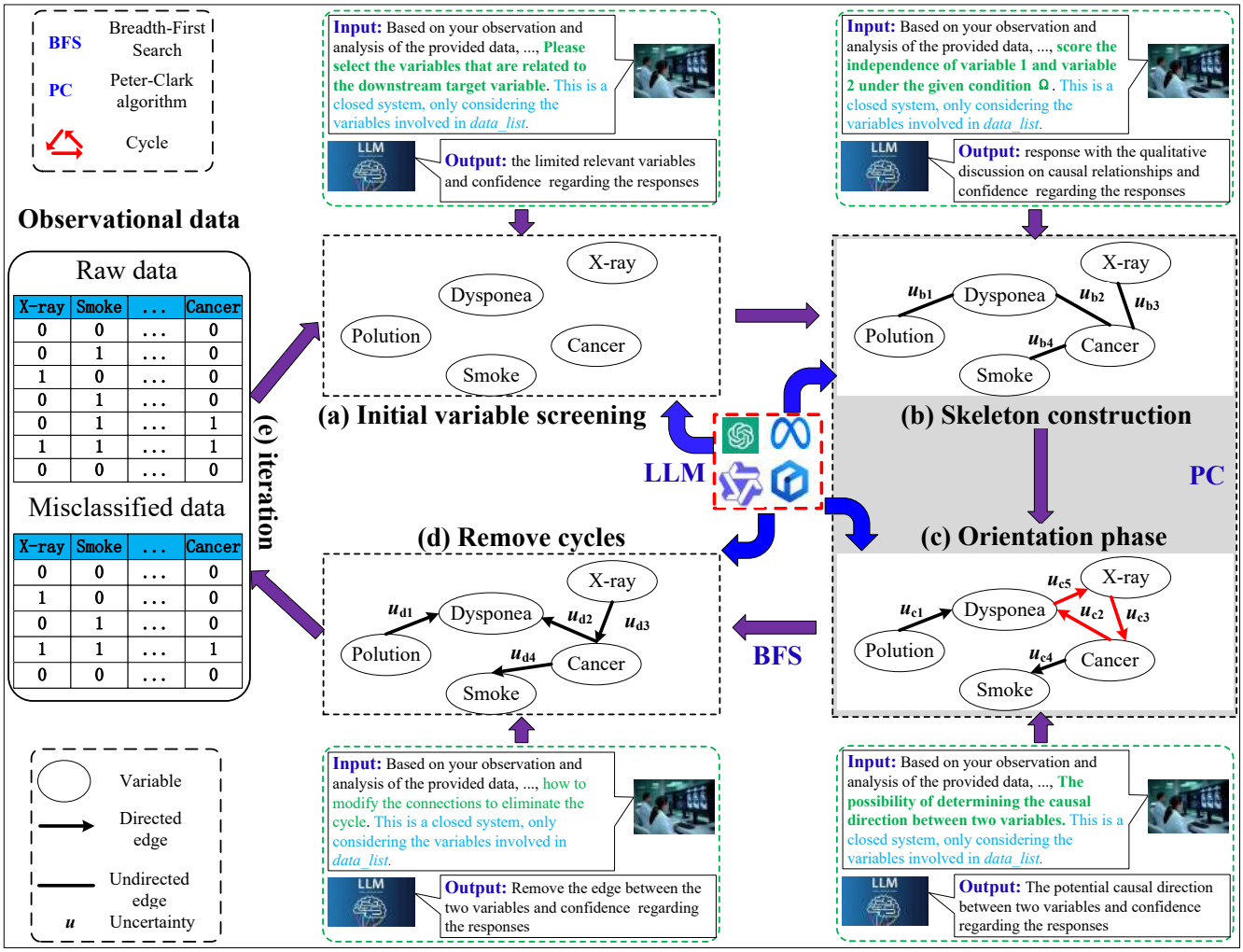


Figure 2: The overview of LLM-CD framework.

and redundant variables. The computation formula is as follows:

$$\mathcal{X}_S, \mathbb{G}^0 = LLM\left(\mathcal{X}_0\right) , \tag{1}$$

where $\mathcal{X}_0$ represents the set of initial variables contained in the observational data without any screening, $\mathcal{X}_S$ is the set of variables after filtering, $\mathbb{G}^0$ is a causal graph with no relationships for $\mathcal{X}_S$, and LLM represents any LLMs.

2. Skeleton construction: We deeply integrate the general PC algorithm with LLMs, incorporating the prior knowledge of LLMs at different stages of the PC algorithm for causal discovery. Specifically, in the first stage, by considering the conditional independence tests with LLMs prior knowledge, we obtain a more accurate skeleton structure. In the second stage, LLMs are utilized for edge removal or orientation, helping orient undirected edges in the Markov equivalence class and resulting in a more precise DAG.

The skeleton construction phase considers conditional independence between vertices $u, v \in \mathcal{X}_S$, $u \neq v$. And the goal is to make amendments to the skeleton structure, denoted as $\mathcal{A}_{ske}$:

$$\mathcal{A}_{ske}\left(u,v\right) = \begin{cases} 0, & if\ |p-\alpha| < \sigma \cap LLM\left(u,v,\Omega\right) > 0.5 \\ 0, & if p-\alpha \geqslant \sigma \\ 1, & \text{else,} \end{cases} \tag{2}$$

where p represents the probability of conditional independence between variable u and variable v given a specific set Ω, α is the significance level, σ is the threshold, and $LLM\left(u,v,\Omega\right)$ denotes the conditional independence probability. Here, $\mathcal{A}_{ske}\left(u,v\right) = 0$ indicates that the variables u and v are d-separated given set Ω.

3. Edge orientation: In the PC orientation phase, two different cases between paired nodes are considered: directional edge and undirectional edge. Given a pair of vertices $u, v \in \mathcal{X}_S$, $u \neq v$, amendments $\mathcal{A}_{loc}$ are made according to the following formula:

$$\mathcal{A}_{loc}\left(u,v\right) = \begin{cases} \text{ADJUST0}\left(u,v,LLM\right), & \text{if directional edge exists between } u \text{ and } v \\ \text{ADJUST1}\left(u,v,LLM\right), & \text{if undirectional edge exists between } u \text{ and } v \end{cases} \tag{3}$$

where $\text{ADJUST0}\left(u,v,LLM\right)$ is an LLM-based function that outputs one of three actions: {keep, remove, reverse}, $\text{ADJUST1}\left(u,v,LLM\right)$ outputs one of the following options: $\{u \rightarrow v, v \rightarrow u\}$. Please note that we do not consider the case where there are no edges between

paired nodes. The main reasons are as follows: *First*, in this case, there is no theoretical guarantee, and relying solely on the causal relationships inferred from LLMs would have low credibility, leading to many incorrect edges [55]. *Second*, cost and efficiency should be considered given the combinatoric search space of all possible edges [20].

4. Remove cycles: After the previous steps, the obtained causal graph may contain cyclic structures. To obtain a DAG, existing studies utilize LLMs to review the causal graph from a global perspective and remove cycles [1]. However, existing LLMs are not capable of accurately identifying and directly eliminating cyclic structures [6]. Therefore, we combine traditional cycle detection algorithms with the LLMs, first using BFS to detect cyclic structures, and then employing the LLMs to eliminate the identified cycles. Thus, this process can be formalized as:

$$\mathcal{G}^t = LLM\left(BFS\left(\mathbb{G}^t\right)\right) , \tag{4}$$

where $\mathbb{G}^t$ is the causal graph obtained through the above process. And $\mathcal{G}^t$ refers to the DAG discovered at round t.

5. Iteration: Due to the hallucination of large models and the inability of LLMs to determine complex causal relationships between variables in a single step, iterative assessments are required. Unlike previous studies [29], we focus only on the parent nodes[2] within the Markov blanket of the target node y, which are denoted as FMB (y). First, suppose there exists some variable $w \in \text{FMB}(y)$, but it has not been discovered as parent node of y (i.e., $w \notin \text{FMB}^{\leqslant t}(y)$, where $\text{FMB}^{\leqslant t}(y)$ denote the union of the parent nodes of the target node y discovered in the first t rounds. Then, by the property of Markov blanket $\text{MB}^t(y)$, we know that w is not conditionally independent of y [37]. Therefore, we can derive the following property:

$$H\left(y \mid \text{FMB}^{\leqslant t}(y), w\right) < H\left(y \mid \text{FMB}^{\leqslant t}(y)\right) , \tag{5}$$

where $H(\cdot)$ refers to the entropy and t is the number of iterations. This also means incorporating w into the input of downstream prediction models can facilitate the explanation and prediction of y. If *LLM* can not find the desired w, it means $\text{FMB}^{\leqslant t}(y)$ is sufficient to capture the main information in y. When given the $\mathcal{G}^t$, *LLM* is expected to find useful $\widehat{w}$ such that:

$$H_{\mathcal{G}^t}\left(y \mid \text{FMB}^{\leqslant t}(y)\right) - H_{\mathcal{G}^t}\left(y \mid \text{FMB}^{\leqslant t}(y), \widehat{w}\right) \geqslant 0 , \tag{6}$$

where $H_{\mathcal{G}^t}\left(y \mid \text{FMB}^{\leqslant t}(y)\right)$ refers to the conditional entropy measured on $\mathcal{G}^t$. If the observational data contains sufficiently diverse examples, and the LLMs are sufficiently powerful, Eq.5 can help progressively uncover all the parent nodes in FMB (y). Therefore, to identify the desired variables, we are motivated to generatre an appropriate $\mathcal{G}^{t+1}$ for the next iteration such that:

$$\mathcal{G}^{t+1} = \arg\max_{\widehat{\mathcal{G}} \in \mathcal{G}'} H_{\widehat{\mathcal{G}}}\left(y \mid \text{FMB}^{\leqslant t}(y)\right) , \tag{7}$$

which means the $\text{FMB}^{\leqslant t}(y)$ cannot adequately explain the target variable y. And $\mathcal{G}'$ refers to the set of all possible DAG that can be constructed based on all samples. Recall that one of the key metrics for assessing the quality of the discovered Markov blanket variables is predictivity [37]. Therefore, Eq. 7 can be solved by converting it into a classification problem, where $\mathcal{G}^{t+1}$ represents the DAG constructed from samples that the downstream classification model failed to classify correctly. In our experiments, we perform sample selection for $\widehat{D}_t$ based on the classification with respect to $\text{FMB}^{\leqslant t}(y)$ using the following expression:

$$\widehat{D}_t = f_{\text{miss}}\left(\mathcal{D}, \text{FMB}^{\leqslant t}(y)\right) , \tag{8}$$

where $\mathcal{D}$ represents the set of samples, f_{miss} is the classification model, and $\widehat{D}_t$ is the input sample for the next iteration. After multiple iterations, the final DAG $\mathcal{G}^*$ can be obtained:

$$\mathcal{G}^* = \mathcal{G}^1 \cup \mathcal{G}^2 \cup, \cdots, \mathcal{G}^t . \tag{9}$$

Integrating the above steps, we establish the overall learning procedure to obtain more accurate causal graphs.

3.3 Uncertainty Analysis

This section introduces the theory of Evidence Deep Learning (EDL) [2, 40] to thoroughly analyze the uncertainty generated by LLM-CD. First, we present the relevant knowledge of EDL, and then, in conjunction with our method, analyze the uncertainty under different scenarios.

3.3.1 Evidential Deep Learning. Existing LLMs typically use softmax on top of deep neural networks (DNN) to predict the next token [42]. However, the softmax function has significant limitations in the following aspects [11, 55]. First, the predicted class probabilities are compressed by the denominator of softmax, which leads to overconfident predictions for unseen data. This is particularly detrimental in classification tasks for complex problems. Second, the output of softmax is essentially a point estimate of the multinomial distribution over class probabilities, meaning that softmax cannot capture the uncertainty of class probabilities.

To overcome these limitations, recent EDL was developed to overcome the limitations of softmax-based DNNs by introducing the evidence framework of Dempster-Shafer Theory (DST) [41] and the subjective logic (SL) [22]. EDL provides a principled way to jointly formulate the multiclass classification and uncertainty modeling. In particular, given a sample $\mathbf{x}_{(i)}$ for K-class classification, assuming that class probability follows a prior Dirichlet distribution, the negative log-likelihood loss to be minimized for learning evidence $\mathbf{e}_{(i)} \in \mathbb{R}_+^K$ eventually reduces to the following form:

$$\begin{aligned} \mathcal{L}_{edl,i}(\mathrm{y}, e; \theta) &= -\log\left(\int \prod_{k=1}^{K} p_{ik}^{y_{ik}} \frac{1}{B(\alpha_i)} \prod_{k=1}^{K} p_{ik}^{\alpha_{ik}-1} d\mathbf{P}_i\right) \\ &= \textstyle\sum_{k=1}^{K} y_{ik}\left(\log(\mathcal{S}_i) - \log(e_{ik}+1)\right) , \end{aligned} \tag{10}$$

where $\mathbf{y}_i$ is an one-hot K-dimensional label for sample $\mathbf{x}_{(i)}$ and $\mathbf{e}_{(i)}$ can be expressed as $\mathbf{e}_{(i)} = g\left(f\left(\mathbf{x}_{(i)}; \theta\right)\right)$. Here, f is the output of a DNN parameterized by θ and g is the evidence function to keep evidence $\mathbf{e}_k$ non-negative. $B(\alpha_i)$ represents the K-dimensional multinomial beta function. $\mathbf{P}_i$ is a simplex representing class assignment probabilities. $\mathcal{S}_i$ is the total strength of a Dirichlet distribution $D(\mathbf{p} \mid \alpha)$, which is parameterized by $\alpha \in \mathbb{R}^K$, and $\mathcal{S}$ is defined as $\mathcal{S} = \sum_{k=1}^{K} \alpha_k$. Based on DST and SL theory, the α_k is linked to the learned evidence $\mathbf{e}_k$ by the equality $\alpha_k = \mathbf{e}_k + 1$. In the inference phase, the predicted probability of the k-th class is $\hat{\mathbf{p}}_k = \alpha_k / \mathcal{S}$

[2]The reason we do not use the Markov blanket of the target variable here is twofold: first, we are more interested in obtaining the parent nodes of the target variable, rather than its child nodes or the parents of nodes that share the same child nodes as the target variable; second, as demonstrated by the subsequent experiments, the parent nodes of the target variable are more predictive.

and the predictive uncertainty u can be deterministically given as $u = K/\mathcal{S}$. For more details, please refer to [40].

3.3.2 Uncertainty Estimation in LLM-CD. To fully explore the uncertainty of the model, we discuss two scenarios: the white-box model (such as the LLAMA series) and the black-box model (such as the GPT series).

1. For white-box models: First, we assume that when generating each token[3], the probability vector $\mathbf{p} \in \mathbb{R}^K$ output by the LLM follows a Dirichlet distribution [36], i.e., $D(\mathbf{p} \mid \alpha)$. This assumption is reasonable and consistent with [4]. Next, we transform the instruction set during fine-tuning into a classification problem.

For pre-training or fine-tuning stages: To enhance the reasoning capabilities of LLM-CD using benchmark datasets with causal edge information between pairwise variables, we replace the final softmax layer of the LLMs with another activation function, such as ReLU, to ensure non-negative outputs. These outputs are then used as the evidence vector for predicting the Dirichlet distribution. In this case, the following equation (with the detailed derivation provided in Appendix B.1) can serve as the objective function for model training, enabling the model to output the calibrated predicted probabilities and their corresponding uncertainty values:

$$\begin{aligned}\mathcal{L}(\theta) = &\sum_{i=1}^{m} \textstyle\sum_{k=1}^{K} y_{ik} \left(\log\left(\mathcal{S}_i\right) - \log\left(e_{ik}+1\right)\right) \\ &+ \lambda_t \sum_{i=1}^{m} KL\left[\mathrm{Dir}\left(\mathbf{p_i} \mid \tilde{\alpha}_{x_i}\right) \parallel \mathrm{Dir}\left(\mathbf{p_i} \mid \langle 1, \cdots, 1\rangle\right)\right],\end{aligned} \tag{11}$$

where $\lambda_t = \min(1.0, t/10) \in [0,1]$ is the annealing coefficient, and t is the index of the current training epoch, m represents the number of queries made to LLMs in LLM-CD, and KL represents the KL divergence.

For inference stages: In inference, our method only requires a single forward pass to calculate the uncertainty. We average the evidence vectors obtained multiple times, and then compute the uncertainty for the answer corresponding to each query according to the following formula (the detailed derivation can be found in Appendix B.2):

$$u = \frac{K_i}{\sum_{l=1}^{K_i} \sigma\left(f\left(x_i\right)\right)_l + 1}, \tag{12}$$

where K_i denotes the number of labels corresponding to query x_i [4], σ is an activation function other than softmax, and $f(x_i)$ refers to the white-box LLMs.

2. For black-box models: Unlike white-box models, black-box models do not allow direct access to the internal structure of the model. Existing research typically only considers utilizing the model's output confidence or training an additional neural network to measure uncertainty [42, 55]. Unlike previous studies, considering the confidence output from LLMs and the distribution differences between responses, we designed a novel uncertainty metric specifically for black-box LLMs. The specific calculation formula is as follows:

$$\widehat{u} = \sum_{j=1}^{n} \left(\frac{C}{KL\left(\bar{\phi}, \phi_j\right)}\right), \tag{13}$$

where C is the confidence of the LLMs output, $\bar{\phi}$ is the mean of the probability distribution, ϕ_j is the probability distribution corresponding to each response, such as [0.8, 0.08, 0.12] for a certain directed edge, and n represents the number of responses corresponding to a query. The larger the $\widehat{u}$, the more stable the output.

[3]For example, an answer to a query in the fine-tuning instruction set is: yes.

[4]One inquiry is one sample, and we design the causal question-answering instructions as a classification task.

Table 1: Statistics of datasets.

Dataset	#Domain	#Sample	#The number of variables	#Labels
WCHSU	Medical	200,000	16/51	2
MIMIC-IV	Medical	4,630	18	2
Asia	Social Science	10,000	8	2
Child	Social Science	10,000	20	2

4 EXPERIMENTS

Our investigation focuses on addressing the following research questions: **RQ1**: How does the performance of LLM-CD compare with that of existing methods? **RQ2**: What is the impact of each component in LLM-CD on the overall performance? **RQ3**: How do human experts evaluate the outputs of LLM-CD? **RQ4**: What are the behavioral patterns of LLM-CD? **RQ5**: What is the uncertainty of LLM-CD? **RQ6**: What are the costs involved? **RQ7**: How does RTGCN respond to alterations in hyperparameter settings? **RQ8**: How does LLM-CD perform in real-world cases?

4.1 Experimental Setup

4.1.1 Datasets. In this section, we evaluate our method on four real-world datasets, including two medical domain datasets and two generic domain benchmark datasets. As shown in Table 1, we utilized the de-identified WCHSU and MIMIC-IV [21] datasets from real hospitals scenarios. Additionally, we also use the classic benchmark datasets in the Bayesian networks literature: Asia [25] and Child [43], both with a sample size of n = 10,000. The metadata descriptions of the random variables are adapted from [31]. Due to the limitations of the PC algorithm, we are not currently addressing cycles and latent confounder, which will be the focus of our future research. Further data details are supplied in the Appendix A.2.

4.1.2 Baselines. We compared three types of baselines: TCDA, LLMs, and LLM-based hybrid methods. The TCDA include PC and FCI, while the LLMs include GPT-3.5, GPT-4, GPT-4O, GPT-4O-mini[5], and LLAMA-3[13]. Additionally, the hybrid methods include LLM-BFS [20], LLM-greedy[31], and ChatPC [8].

4.2 RQ1: Main results

Consistent with recent works [29, 35], downstream classification tasks are regarded as an effective approximate method for evaluating causal graphs, especially when real causal graphs are unavailable. Other recent studies [14] have also shown that uncovering the correct causal structure can improve predictive performance on target variables. In addition, we have also conducted detailed human evaluations and case studies to further assess the constructed causal graphs. Please note that this paper does not address the handling of

[5]https://openai.com/api/.

Table 2: Performance evaluation for lung cancer prediction tasks using GPT-4. OOCL indicates that the model has exceeded its maximum context length. ZSL = Zero-Shot Learning; FSL = Few-Shot Learning.

Methods	WCHSU (n = 16)				WCHSU (n = 51)				MIMIC-IV			
	ACC	Recall	Percision	AUC	ACC	Recall	Percision	AUC	ACC	Recall	Percision	AUC
PC	0.9929±0.0147	0.0070±0.0024	0.5000±0.0036	0.5225±0.0017	0.9651±0.0124	0.0248±0.0132	0.0093±0.0021	0.5903±0.0037	0.8734±0.0056	0.8120±0.0103	0.6368±0.0025	0.5349±0.0026
FCI	0.9929±0.0113	0.0070±0.0017	0.5000±0.0147	0.5247±0.0158	0.9651±0.0203	0.0248±0.0034	0.0093±0.0017	0.6119±0.0216	0.5583±0.0041	1.0000±0.0013	0.5272±0.0007	0.5500±0.0139
GPT-3.5	0.9320±0.0168	0.0944±0.0053	0.0108±0.0035	0.4966±0.0114	0.9205±0.0176	0.1224±0.0194	0.0118±0.0016	0.4986±0.0098	0.5248±0.0062	1.0000±0.0015	0.5089±0.0054	0.5508±0.0192
GPT-4	0.9104±0.0113	0.1783±0.0248	0.0150±0.0020	0.5274±0.0206	0.7811±0.0169	0.5070±0.0243	0.0166±0.0009	0.4994±0.0114	0.5583±0.0062	1.0000±0.0019	0.5272±0.0027	0.5319±0.0147
GPT-4O	0.9104±0.0191	0.1783±0.0183	0.0150±0.0023	0.5280±0.0166	0.9790±0.0213	0.0169±0.0029	0.0071±0.0007	0.6450±0.0213	0.5572±0.0081	1.0000±0.0031	0.5266±0.0104	0.5201±0.0124
GPT-4O-mini	0.9105±0.0215	0.1783±0.0418	0.0150±0.0059	0.5271±0.0183	0.9790±0.0284	0.0169±0.0061	0.0071±0.0024	0.6450±0.0173	0.5583±0.0118	1.0000±0.0016	0.5272±0.0159	0.5691±0.0192
LLM-BFS	0.9853±0.0304	0.0114±0.0046	0.0058±0.0015	0.5462±0.0236	OOCL	OOCL	OOCL	OOCL	0.5583±0.0063	1.0000±0.0072	0.5272±0.0031	0.6043±0.0186
ChatPC	0.9769±0.0138	0.0325±0.0121	0.0164±0.0026	0.5396±0.0095	0.8729±0.0318	0.3527±0.0163	0.0169±0.0064	0.6419±0.0115	0.5472±0.0053	0.9805±0.0162	0.5301±0.0043	0.6127±0.0181
LLM-greedy	0.9929±0.0196	0.0070±0.0014	0.5000±0.0095	0.5427±0.0106	0.8920±0.0277	0.2238±0.0304	0.0154±0.0061	0.6532±0.0134	0.5583±0.0058	1.0000±0.0113	0.5272±0.0046	0.6079±0.0158
LLM-CD (ZSL)	0.0780±0.0141	0.8811±0.0376	0.0068±0.0010	0.5873±0.0146	0.6296±0.0156	0.6958±0.0176	0.0133±0.0007	0.6947±0.0141	0.5731±0.0022	1.0000±0.0036	0.5384±0.0011	0.6451±0.0169
LLM-CD (FSL, k=2)	0.0654±0.0176	**0.8985±0.0395**	0.0064±0.0012	**0.6115±0.0106**	0.6279±0.0172	**0.7218±0.0364**	0.0128±0.0003	**0.7126±0.0089**	0.5601±0.0003	1.0000±0.0052	0.5291±0.0005	**0.6566±0.0118**

Table 3: Performance comparison on benchmark datasets.

Methods	Asia					Child				
	ACC	Recall	Percision	NPE	Ratio ↓	ACC	Recall	Percision	NPE	Ratio ↓
PC	0.9412±0.0094	0.0930±0.0085	0.9412±0.0094	6±0.9336	0.2857±0.0126	0.7480±0.0112	1.0000±0.0091	0.7480±0.0112	25±1.0242	0.2400±0.0218
FCI	0.9390±0.0142	0.1000±0.0079	0.9390±0.0142	8±0.8283	0.3225±0.0532	0.7480±0.0237	1.0000±0.0081	0.7480±0.0237	28±0.9223	0.4124±0.0716
GPT-3.5	0.9390±0.0106	1.0000±0.0082	0.9390±0.0106	7±0.9457	0.1667±0.0710	0.8915±0.0114	0.7167±0.0047	0.8915±0.0114	15±1.4306	0.7000±0.0584
GPT-4	0.9390±0.0083	1.0000±0.0075	0.9390±0.0083	8±0.9173	0.2132±0.0817	0.7480±0.0168	1.0000±0.0043	0.7480±0.0168	14±1.1134	0.6923±0.0957
GPT-4O	0.9390±0.0153	1.0000±0.0085	0.9390±0.0153	8±1.0425	0.1925±0.0537	0.7480±0.0172	1.0000±0.0103	0.7480±0.0172	13±0.9564	0.6316±0.0742
GPT-4O-mini	0.9390±0.0184	1.0000±0.0126	0.9390±0.0184	14±1.2315	0.4545±0.0336	0.7480±0.0173	1.0000±0.0149	0.7480±0.0173	36±2.0105	0.9344±0.0849
LLM-BFS	0.9440±0.0256	1.0000±0.0189	0.9440±0.0256	15±1.4171	0.3913±0.0337	0.7325±0.0433	1.0000±0.0182	0.7325±0.0433	11±1.5435	0.8333±0.0742
ChatPC	0.9402±0.0203	0.2166±0.0101	0.9402±0.0203	17±0.9986	0.6473±0.0184	0.7681±0.0259	0.7248±0.0173	0.7681±0.0259	23±1.3437	0.7157±0.0862
LLM-greedy	0.9713±0.0131	0.0764±0.0215	0.9713±0.0131	11±0.8712	0.8947±0.0224	0.7480±0.0185	1.0000±0.0089	0.7480±0.0185	36±1.6517	0.8689±0.1384
LLM-CD (ZSL)	0.9390±0.0051	1.0000±0.0178	0.9390±0.0051	6±1.0218	**0.1429±0.0673**	0.7385±0.0064	1.0000±0.0146	0.7385±0.0064	16±0.5869	0.2674±0.0947
LLM-CD (FSL, k=2)	0.9430±0.0054	1.0000±0.0155	0.9430±0.0054	6±0.8164	0.1857±0.0639	0.7480±0.0083	1.0000±0.0232	0.7480±0.0083	15±1.0316	**0.2100±0.1012**

spurious features. Although this is undoubtedly a valuable future research direction, it is beyond the scope of this work.

4.2.1 Experiments on medical datasets. **Experimental setup:** The WCHSU and MIMIC-IV datasets are divided into training and test sets in an 8:2 ratio. For the WCHSU, consistent with existing research [35, 48], we set the variable *Lung Cancer* as the target node. Based on the causal graphs obtained using different methods, we identified the parent nodes of the target node and then made predictions using the variable values of these parent nodes. A similar setup was followed for the MIMIC-IV data. ACC, Recall, AUC, and Precision were used as evaluation metrics. Please note that in this real-world scenario, our goal is to correctly identify early-stage lung cancer patients, so we focus more on the recall metric, which represents the ability to correctly identify positive samples.

Results: As shown in Table 2, LLM-CD outperforms the second-best model by an average of 169.53% in the Recall metric and 9.92% in AUC, with the highest improvement reaching up to 403.93% on the WUSCH (n = 16) dataset. Specifically, TCDA such as PC and FCI, despite having high ACC and precision, are unable to correctly classify lung cancer patients based on the DAG they generate. These methods are mainly data-driven, and the DAG they produce are susceptible to the influence of the data itself. Knowledge-driven large model approaches utilize prior knowledge for causal discovery, and compared to TCDA, they exhibit some causal reasoning capabilities, which is consistent with existing research [17, 23]. The latest LLMs-based baselines, although incorporating the prior knowledge of LLMs, fail to fully integrated with TCDA, thereby resulting in suboptimal performance.

4.2.2 Experiments on benchmark datasets. **Experimental setup:** The Asia and Child datasets are divided into training and testing sets in an 8:2 ratio. For the Asia dataset, we select the variable *Lung Cancer* as the target node. Based on the causal graphs obtained through different methods, we identify the parent nodes of the target node and use the variable values of these parent nodes for prediction. Consistent with previous studies [34], we use the variable *GruntingReport* in the Child dataset as the target node and adopt a similar setup to that of the Asia dataset. Since the benchmark datasets contain ground truth causal graphs, in addition to performance metrics (ACC, Recall, Percision) for downstream tasks, we also use NPE and Ratio[6].

Results: As shown in Table 3, we can observe that the prediction performance based on the target node is similar across different methods, with no significant differences. This may be related to the data sample size and the nature of the selected target node. On Recall and Ratio, our proposed method demonstrates a significant advantage, with improvements of up to 25.77% . GPT-4 and GPT-4O perform the second best on the benchmark dataset, potentially because their pre-training data contain relevant knowledge. GPT-4O-mini, however, performs worse than TCDA.

4.3 RQ2: Ablation Studies

To evaluate the effectiveness of different components in LLM-CD, we conduct the ablation study with several variants which are

[6]NPE: the number of predicted edges; Ratio: the ratio between Normalized Hamming Distance (NHD) and baseline NHD, and the smaller the Ratio, the more accurate the discovered causal graph is. The NHD refers to the number of edges that are present in one graph but not the other, divided by the total number of all possible edges.

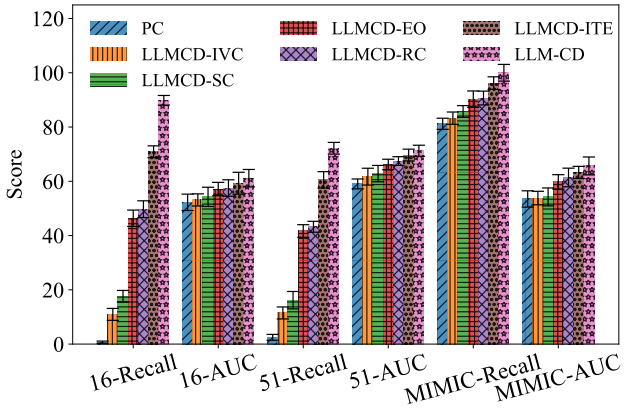


Figure 3: Performance of evaluation for ablation study on WCHSU. And 16 and 51 represent the number of variables.

Table 4: Comparison of performance using different LLMs.

Methods	WCHSU ($n = 16$) ACC	Recall	Percision	AUC	WCHSU ($n = 51$) ACC	Recall	Percision	AUC
GPT-3.5	0.0910	0.8601	0.0067	0.5964	0.6317	0.6923	0.0133	0.6714
GPT-4	0.0637	0.8671	0.0066	0.5905	0.6296	0.6958	0.0133	0.6839
GPT-4O	0.0805	0.8916	0.0069	0.5986	0.6312	0.6923	0.0133	0.6832
GPT-4O-mini	0.0578	0.9231	0.0070	0.6273	0.7723	0.3427	0.0092	0.3356
DeepSeek-v3	0.0837	0.8322	0.0065	0.5667	0.0493	0.9091	0.0068	0.8995
LLAMA-3.1-70B	0.0681	0.9126	0.0070	0.6231	0.1091	0.8957	0.0089	0.8830
LLAMA-3.1-8B	0.1837	0.7452	0.1316	0.5172	0.7916	0.2148	0.0159	0.2021

introduced as follows: 1) **LLMCD-IVS**: integrates LLMs into the Initial variable screening (IVS). 2) **LLMCD-SC**: integrates the LLMs into the IVS and skeleton construction stage (SC). 3) **LLMCD-EO**: Integrates LLMs into the IVS, SC, and edge orientation (EO). 4) **LLMCD-RC**: Does not include iterative refinement (IR) or uncertainty estimation (UE) components. 5) **LLMCD-ITE**: Excludes the UE component. Additionally, we analyzed the effects of different LLMs performance.

As shown in Figure 3 and Table 4 , we have the following observations: 1) LLM-CD performs best when all components are included. However, the performance difference on the other two metrics is not as pronounced. 2) When the IE and UE modules are added, LLM-CD shows the greatest performance improvement, highlighting their important impact on model performance. 3) There are noticeable performance differences across different LLMs, with GPT-4O-mini and DeepSeek-v3 showing relatively better performance.

4.4 RQ3: Human Evaluations

To further verify the validity of the DAG generated by our method, we invited ten experts from the largest hospital in Asia in the relevant field to rate the DAG. The detailed process can be found in Appendix C.1. Each edge was scored on a scale of 1 to 10, with higher scores indicating a higher probability that the causal edge was correctly identified. Edges with average scores between 1 and 3 were considered non-existent, edges with average scores between 4 and 6 are considered uncertain about their existence, and edges with scores between 7 and 10 were considered correctly identified. Then, we calculated the corresponding proportions. As shown in Table 5, we can see that LLM-CD is able to generate more accurate DAG, thereby obtaining more valuable insights.

Table 5: Human Evaluation for WCHSU and MIMIC data. Due to time and cost considerations, we only compared the results of the well-performing baseline according to Table 2.

Methods	Type	WCHSU ($n = 16$) Number of edges	Ratio	MIMIC-IV Number of edges	Ratio
GPT-4	Correct	6	31.58	7	38.89
	Incorrect	8	42.11	5	27.78
	Uncertain	5	26.31	6	33.33
LLM-CD	Correct	12	**70.6**	11	**68.8**
	Incorrect	3	**17.6**	2	**12.5**
	Uncertain	2	**11.8**	3	**18.7**

4.5 RQ4: LLM Behavioural Experiments

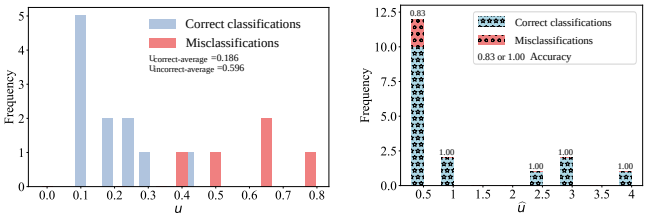
Early empirical studies have shown that LLMs can significantly reduce the size of the Markov equivalence class of a given DAG and have a high probability of retaining the ground-truth causal graph [1, 31]. To better understand this phenomenon, we investigate the behavior patterns of LLM-CD in reasoning over edges in the causal graph space. Figure 8 in Appendix C.2 illustrates the modification of hypotheses across two different types of relationships. We first analyze the causal relationship between Disease and Duct-flow as an example. If the starting point is an undirected edge relationship, LLM-CD most often tends to output the edge Duct-flow→Disease. When an edge already exists in this direction (as shown in Figure 8 in Appendix C.2, the first two subfigures), the same conclusion is drawn. Overall, LLM-CD prefers the causal relationship Duct-flow→Disease but also considers the reverse relationship Disease→Duct-flow, especially when the initial direction is uncertain. This is consistent with the idea that disease processes can sometimes lead to secondary effects, such as the disruption of duct flow, particularly in complex medical conditions where feedback loops or compensatory mechanisms might occur. More analysis can be found in Appendix C.2.

4.6 RQ5: Uncertainty Estimation in LLM-CD

We primarily explore the uncertainty estimation of LLM-CD from both open-source and closed-source perspectives. For the closed-source black-box models, we take the GPT-4O as an example and use Eq. 13 to quantify the uncertainty of LLM-CD. For the open-source white-box models, we use LLAMA-3.1-8B as an example and measure uncertainty using Eq. 12. We use the Child dataset as an example for analysis, and Figure 4 shows the specific distributions. As shown in Figure 4(a), edges with correct classifications exhibit low uncertainty, while LLM-CD assigns higher uncertainty values to edges with incorrect classifications. In Figure 4(b), we can see that when $\widehat{u}$ exceeds 0.49, the model provides more accurate answers. Furthermore, inspired by existing research [18], we explore the impact of 20 causally related vocabulary terms on our model using the Child dataset. As shown in Table 6, our method consistently achieves good experimental results under different prompt words.

4.7 RQ6: Cost Analysis

We reported the costs of our method at different stages and with varying sample sizes, as shown in Figure 5. It is evident that the costs differ across various LLMs: GPT-4O-mini has the lowest cost but the worst performance, while DeepSeek demonstrates competitive performance with moderate costs. GPT-4O has the highest cost but



(a) LLAMA-3.1-8B (b) GPT-4O

Figure 4: The uncertainty distribution across different LLMs.

Table 6: Model performance under different phrases.

Phrases	ACC	Child Recall	Child Percision	NPE	Ratio ↓
Provokes	0.7515	1.0000	0.7515	20	0.7778
Triggers	0.7235	1.0000	0.7235	19	0.6818
Leads to	0.7401	1.0000	0.7401	16	0.3218
Induces	0.7380	1.0000	0.7380	21	0.6087
Results in	0.7345	1.0000	0.7345	21	0.6087
Brings about	0.7460	1.0000	0.7460	16	0.3018
Yields	0.7560	1.0000	0.7560	15	0.3018
Generates	0.7395	1.0000	0.7395	17	0.3778
Initiates	0.7430	1.0000	0.7430	17	0.3273
Produces	0.7550	1.0000	0.7550	15	0.2878
Stimulates	0.7465	1.0000	0.7465	18	0.3209
Instigates	0.7505	1.0000	0.7505	17	0.2696
Fosters	0.7361	1.0000	0.7361	19	0.3696
Engenders	0.7470	1.0000	0.7470	20	0.3878
Promotes	0.7255	1.0000	0.7255	20	0.4878
Catalyzes	0.7430	1.0000	0.7430	20	0.4681
Gives rise to	0.7455	1.0000	0.7455	18	0.3482
Spurs	0.7523	1.0000	0.7523	21	0.4087
Sparks	0.7305	1.0000	0.7305	18	0.2737
Causes	0.7480	1.0000	0.7480	15	0.2100
Average	0.7426±*0.0089*	1.0000±*0.0000*	0.7480±*0.0089*	18.15±*2.0316*	0.4415±*0.1442*

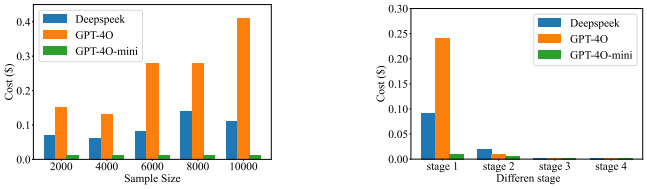


Figure 5: The cost statistics of causal discovery using our model with different LLMs on Child datasets.

the best performance. Therefore, DeepSeek might be the optimal choice for our task. Moreover, the costs produced by our method do not exhibit a linear relationship with increasing sample sizes, indicating good economic efficiency. Regarding the costs at different stages, they align with real-world expectations, as the causal graph generated shows that the first and second stages correspond to the most edges, while undirected edges and cycles are relatively rare.

4.8 RQ7: Parameter Sensitivity

We further investigate the effect of model parameters on downstream prediction tasks. In the Child datasets, we explore the impact of the threshold σ, significance level α, temperature coefficient τ, and uncertainty threshold ω. The results are shown in Figure 6 and Figure 10 in Appendix C.4. Overall, compared to α, the σ and ω have a greater impact on model performance. As the σ increases, the performance first improves and then decreases, with the optimal

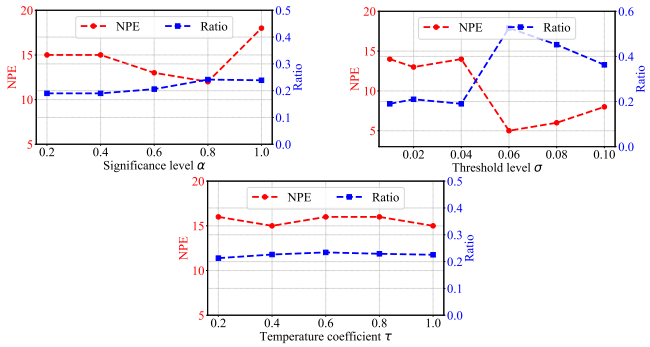
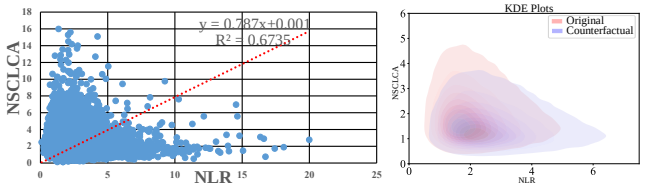


Figure 6: Performance on Child with different α, σ, and τ.



(a) Statistical distribution (b) Counterfactual distribution

Figure 7: The statistical distribution and counterfactual distribution of NLR and NSCLC.

value ranging from 0 to 0.08. However, we found that the temperature coefficient τ has a minimal impact on the model's performance. This could be because we are outputting probability values for each query rather than generating complete text segments.

4.9 RQ8: Case Studies

Using the WCHSU dataset as an example, LLM-CD identifies a high probability that Neutrophil-to-Lymphocyte Ratio (NLR) has a direct effect on Non-small cell lung cancer antigen level (NSCLC). First, The statistical observation of the two variables shows that there is a correlation between them (as shown in Figure 7a). However, to investigate this relationship from a causal perspective, we performed an intervention on NLR using counterfactual inference, and the resulting kernel density estimation curve is shown in Figure 7b. The results indicate that through counterfactual inference, an increase in NLR leads to an increase in the measurement of NSCLC. There is a borderline statistically significant difference between the observational and counterfactual distributions. We believe this likely represents a missing edge in the ground-truth causal graph provided by [30, 56].

5 CONCLUSION

In this paper, we propose a new framework named LLM-CD to integrate the rich knowledge of LLMs into the causal discovery process. Specifically, LLM-CD deeply incorporates the reasoning capabilities of LLMs at various stages of causal discovery. Then, an iterative process is designed to further enhance the causal discovery potential of LLMs. Subsequently, we leverage the theory of evidence-based deep learning to quantify the uncertainty of our approach, enabling more reliable inference results. Finally, we conduct extensive experiments on a large-scale real-world hospital dataset we constructed,

a dataset based on MIMIC-IV, and two benchmark datasets, validating the effectiveness of the proposed LLM-CD framework in both causal discovery and downstream tasks.

ACKNOWLEDGMENTS

This research is partially supported by funding from Xiangjiang Laboratory (25XJ02002), grant from the National Natural Science Foundation of China (72495125, 62376227, 62376228), the science and technology innovation Program of Hunan Province (2024RC4008, AC2024040911247631ff26), and Sichuan Science and Technology Program (2023NSFSC0032, 2024YFFK0059). Carl Yang and Yujia Zheng are not supported by any funds from China.

REFERENCES

[1] A. Abdulaal, N. Montana-Brown, T. He, A. Ijishakin, I. Drobnjak, D. C. Castro, D. C. Alexander, et al. Causal modelling agents: Causal graph discovery through synergising metadata-and data-driven reasoning. In *ICLR*, 2023.
[2] A. Amini, W. Schwarting, A. Soleimany, and D. Rus. Deep evidential regression. In *NeurIPS*, volume 33, pages 14927–14937, 2020.
[3] T. Ban, L. Chen, D. Lyu, X. Wang, Q. Zhu, and H. Chen. Llm-driven causal discovery via harmonized prior. *TKDE*, 2025.
[4] W. Bao, Q. Yu, and Y. Kong. Evidential deep learning for open set action recognition. In *Proceedings of the ICCV*, pages 13349–13358, 2021.
[5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
[6] S. Chen, M. Xu, K. Wang, X. Zeng, R. Zhao, S. Zhao, and C. Lu. Clear: Can language models really understand causal graphs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6247–6265, 2024.
[7] D. M. Chickering. Optimal structure identification with greedy search. *JMLR*, 3(Nov):507–554, 2002.
[8] K.-H. Cohrs, E. Diaz, V. Sitokonstantinou, G. Varando, and G. Camps-Valls. Large language models for constrained-based causal discovery. In *AAAI 2024 Workshop on"Are Large Language Models Simply Causal Parrots?"*, 2023.
[9] S. Dash, V. N. Balasubramanian, and A. Sharma. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 915–924, 2022.
[10] M. Farooq, S. Hardan, A. Zhumbhayeva, Y. Zheng, P. Nakov, and K. Zhang. Understanding breast cancer survival: Using causality and language models on multi-omics data. *arXiv preprint arXiv:2305.18410*, 2023.
[11] J. Gao, M. Chen, L. Xiang, and C. Xu. A comprehensive survey on evidential deep learning and its applications. *arXiv preprint arXiv:2409.04720*, 2024.
[12] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.
[13] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.
[14] S. Gupta, D. Childers, and Z. C. Lipton. Local causal discovery for estimating causal effects. In *Conference on Causal Learning and Reasoning*, pages 408–447. PMLR, 2023.
[15] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.
[16] Z. Huang, X. Xia, L. Shen, B. Han, M. Gong, C. Gong, and T. Liu. Harnessing out-of-distribution examples via augmenting content and style. In *ICLR*, 2023.
[17] Z. Jin, J. Liu, Z. Lyu, S. Poff, M. Sachan, R. Mihalcea, M. Diab, and B. Schölkopf. Can large language models infer causation from correlation? In *ICLR*, 2023.
[18] Z. Jin, J. Liu, L. Zhiheng, S. Poff, M. Sachan, R. Mihalcea, M. T. Diab, and B. Schölkopf. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*, 2024.
[19] B. Jing, D. Zhou, K. Ren, and C. Yang. Causality-aware spatiotemporal graph neural networks for spatiotemporal time series imputation. In *Proceedings of the CIKM*, pages 1027–1037, 2024.
[20] T. Jiralerspong, X. Chen, Y. More, V. Shah, and Y. Bengio. Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*, 2024.
[21] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.
[22] A. Jøsang. *Subjective logic*, volume 3. Springer, 2016.
[23] E. Kıcıman, R. Ness, A. Sharma, and C. Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv:2305.00050*, 2023.
[24] A. Lampinen, S. Chan, I. Dasgupta, A. Nam, and J. Wang. Passive learning of active causal strategies in agents and language models. *NeurIPS*, 36, 2024.
[25] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.
[26] C. Lee, J. Kim, Y. Jeong, J. Lyu, J. Kim, S. Lee, S. Han, H. Choe, S. Park, W. Lim, et al. Can we utilize pre-trained language models within causal discovery algorithms? *arXiv preprint arXiv:2311.11212*, 2023.
[27] X.-C. Li, K. Zhang, and T. Liu. Causal structure recovery with latent variables under milder distributional and graphical assumptions. In *ICLR*, 2023.
[28] B. Liu, J. T. Ash, S. Goel, A. Krishnamurthy, and C. Zhang. Transformers learn shortcuts to automata. In *ICLR*, 2023.
[29] C. Liu, Y. Chen, T. Liu, M. Gong, J. Cheng, B. Han, and K. Zhang. Discovery of the hidden world with large language models. In *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
[30] J. Liu, S. Li, S. Zhang, Y. Liu, L. Ma, J. Zhu, Y. Xin, Y. Wang, C. Yang, and Y. Cheng. Systemic immune-inflammation index, neutrophil-to-lymphocyte ratio, platelet-to-lymphocyte ratio can predict clinical outcomes in patients with metastatic non-small-cell lung cancer treated with nivolumab. *Journal of clinical laboratory analysis*, 33(8):e22964, 2019.
[31] S. Long, A. Piché, V. Zantedeschi, T. Schuster, and A. Drouin. Causal discovery with language models as imperfect experts. In *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2023.
[32] S. Long, T. Schuster, and A. Piché. Can large language models build causal graphs? In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.
[33] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *JMLR*, 17(32):1–102, 2016.
[34] S. Nadkarni and P. P. Shenoy. A bayesian network approach to making inferences in causal maps. *European Journal of Operational Research*, 128(3):479–498, 2001.
[35] N. Naik, A. Khandelwal, M. Joshi, M. Atre, H. Wright, K. Kannan, S. Hill, G. Mamidipudi, G. Srinivasa, C. Bifulco, et al. Applying large language models for causal structure learning in non small cell lung cancer. In *ICHI*, pages 688–693, 2024.
[36] K. W. Ng, G.-L. Tian, and M.-L. Tang. Dirichlet and related distributions: Theory, methods and applications. 2011.
[37] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.
[38] J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.
[39] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *JMLR*, 15(58):2009–2053, 2014.
[40] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. *NeurIPS*, 31, 2018.
[41] K. Sentz and S. Ferson. *Combination of evidence in Dempster-Shafer theory*. Sandia National Laboratories, 2002.
[42] V. Shrivastava, P. Liang, and A. Kumar. Llamas know what gpts don't show: Surrogate models for confidence estimation. *arXiv:2311.08877*, 2023.
[43] D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, and R. G. Cowell. Bayesian analysis in expert systems. *Statistical science*, pages 219–247, 1993.
[44] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2001.
[45] P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 499–506, 1995.
[46] R. Tu, C. Ma, and C. Zhang. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. *arXiv preprint arXiv:2301.13819*, 2023.
[47] A. Vashishtha, G. R. Abbavaram, A. Kumar, S. Bachu, V. N. Balasubramanian, and A. Sharma. Causal order: The key to leveraging imperfect experts in causal inference. In *ICLR*, 2025.
[48] C. Wang, J. Shao, Y. He, J. Wu, X. Liu, L. Yang, Y. Wei, X. S. Zhou, Y. Zhan, F. Shi, et al. Data-driven risk stratification and precision management of pulmonary nodules detected on chest computed tomography. *Nature Medicine*, pages 1–12, 2024.
[49] X. Wang, T. Ban, L. Chen, D. Lyu, Q. Zhu, and H. Chen. Large-scale hierarchical causal discovery via weak prior knowledge. *TKDE*, 2025.
[50] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
[51] M. Willig, M. Zečević, D. S. Dhami, and K. Kersting. Can foundation models talk causality? In *UAI 2022 Workshop on Causal Representation Learning*, 2022.
[52] M. Willig, M. Zečević, D. S. Dhami, and K. Kersting. Probing for correlations of causal facts: Large language models and causality. *preprint*, 2023.
[53] A. Wu, K. Kuang, M. Zhu, Y. Wang, Y. Zheng, K. Han, B. Li, G. Chen, F. Wu, and K. Zhang. Causality for large language models. *arXiv preprint arXiv:2410.15319*, 2024.
[54] Z. Xie, Y. Zheng, L. Ottens, K. Zhang, C. Kozyrakis, and J. Mace. Cloud atlas: Efficient fault localization for cloud systems using language models and causal insight. *arXiv preprint arXiv:2407.08694*, 2024.

[55] M. Xiong, Z. Hu, X. Lu, Y. LI, J. Fu, J. He, and B. Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*, 2024.

[56] X. Yin, H. Chen, Y. Sun, L. Xiao, H. Lu, W. Guo, H. Yang, J. Zhou, K. Fan, and W. Liang. Prognostic value of neutrophil-to-lymphocyte ratio change in patients with locally advanced non-small cell lung cancer treated with thoracic radiotherapy. *Scientific Reports*, 14(1):11984, 2024.

[57] M. Yue, J. Zhao, M. Zhang, L. Du, and Z. Yao. Large language model cascades with mixture of thought representations for cost-efficient reasoning. In *ICLR*, 2023.

[58] C. Zhang, S. Bauer, P. Bennett, J. Gao, W. Gong, A. Hilmkil, J. Jennings, C. Ma, T. Minka, N. Pawlowski, et al. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*, 2023.

[59] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[60] Y. Zheng, B. Huang, W. Chen, J. Ramsey, M. Gong, R. Cai, S. Shimizu, P. Spirtes, and K. Zhang. Causal-learn: Causal discovery in python. *JMLR*, 25(60):1–8, 2024.

A REPRODUCIBILITY

In this section, we provide more details regarding datasets and experimental setup to facilitate the reproducibility of the results. Our code is available at https://github.com/trytodoit227/LLMCD.

A.1 Datasets

WCHSU. The WCHSU data comes from the health management center of one of the largest hospitals in Asia, consisting of health check-up data from a total of 200,000 participants. The original dataset contains 230 variables. Using LLMs, we scored the correlation between the original variables and the lung cancer variable of interest on a scale of 0 to 5. After filtering based on a score greater than 4, 16 associated variables were selected, while filtering based on a score of 5 resulted in 51 associated variables. The ratio of cancer patients to non-cancer patients in the sample is 147:1.

MIMIC-IV. MIMIC-IV is a publicly available database containing data of patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts, USA. The database includes de-identified data of 383,220 patients admitted to the intensive care unit (ICU) or emergency department (ED) between 2008 and 2019. From this, we have extracted sample data of lung cancer patients. Similar to WCHSU data processing, 16 correlated variables were selected based on a LLMs score of 5.

Since the WCHSU dataset is newly collected and has not been released online, and although the MIMIC-IV is publicly available, policies prohibit its use for training LLMs, especially industrial models like GPT. Therefore, it is reasonable to assume that neither of these datasets has been used for training existing LLMs. Even if some data were leaked into LLM training, any memorized patient records would minimally affect downstream predictions, as only logistic regression is used for prediction, not LLMs.

A.2 Experiment Settings

For the compared methods, we use the source code released by the authors for baseline evaluation. For the LLMs, the temperature coefficient is set to 0.1, and the nucleus sampling method is used with a probability value of 0.7.

In LLM-CD, the significance level α is 0.05, the threshold σ is set to 0.01, the temperature coefficient of the LLMs is set to 0.1, the uncertainty threshold is 0.49, and the kernel sampling method is used with a probability value of 0.7. The downstream classification model is logistic regression, and the number of iterations I is 3.

B UNCERTAINTY ESTIMATION

In this section, we will introduce the detailed derivation process of Eq.11 and Eq. 12.

B.1 Derivation for Eq.11

Inspired by the idea of DEL, we employ a white-box model (such as LLAMA-3.1-8B) in LLM-CD to directly predict the evidence e from the given input x to solve a K-class classification problem [7]. Specifically, the output of our model is activated by a non-negative evidence function. Considering the Dirichlet prior, the LLM-CD is fine-tuned by minimizing the negative log-likelihood (NLL) loss [8]:

$$\begin{aligned}\mathcal{L}_{nll-edl,i}\left(\mathrm{y}, e; \theta\right) &= -\log\left(\int \prod_{k=1}^{K} p_{ik}^{y_{ik}} \frac{1}{B\left(\alpha_i\right)} \prod_{k=1}^{K} p_{ik}^{\alpha_{ik}-1} d\mathbf{P}_i\right) \\ &= -\log\left[\frac{1}{B\left(\alpha_i\right)} \int \prod_{k=1}^{K} p_{ik}^{\alpha_{ik}+y_{ik}-1} d\mathbf{P}_i\right] \\ &= -\log\left[\frac{1}{B\left(\alpha_i\right)} B\left(\alpha_i + \mathrm{y}_i\right)\right] = -\log\left[\frac{B\left(\alpha_i\right)}{B\left(\alpha_i + \mathrm{y}_i\right)}\right] .\end{aligned} \tag{B.1}$$

Also note that:

$$B\left(\alpha_i\right) = \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_{ik}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha_{ik}\right)} = \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_{ik}\right)}{\Gamma\left(\mathcal{S}_i\right)} , \tag{B.2}$$

where $\Gamma(\cdot)$ is the gamma function.

$$B\left(\alpha_i + \mathbf{y}_i\right) = \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_{ik} + y_{ik}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha_{ik} + y_{ik}\right)} = \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_{ik} + y_{ik}\right)}{\Gamma(\mathcal{S}_i+1)} . \tag{B.3}$$

Combining Eq. B.2 and B.3, we obtain:

$$\frac{B(\alpha_i)}{B(\alpha_i+\mathrm{y}_i)} = \frac{\Gamma(\mathcal{S}_i+1)}{\Gamma(\mathcal{S}_i)} \prod_{k=1}^{K} \frac{\Gamma\left(\alpha_{ik}\right)}{\Gamma\left(\alpha_{ik}+1\right)} = \prod_{k=1}^{K} \frac{\mathcal{S}_i}{\alpha_{ik}} = \prod_{k=1}^{K} \left(\frac{\mathcal{S}_i}{\alpha_{ik}}\right)^{y_{ik}} . \tag{B.4}$$

By combining Eq. B.1 and B.4, we get:

$$\begin{aligned}\mathcal{L}_{nll-edl,i}\left(\mathrm{y}, e; \theta\right) &= -\log\left[\frac{B\left(\alpha_i\right)}{B\left(\alpha_i + \mathrm{y}_i\right)}\right] \\ &= \textstyle\sum_{k=1}^{K} y_{ik}\left(\log\left(\mathcal{S}_i\right) - \log\left(e_{ik}+1\right)\right) ,\end{aligned} \tag{B.5}$$

where $\mathrm{y}_i = \{y_{i1}, \cdots, y_{iK}\}$ is an one-hot K-dimensional label for sample x_i and e_i can be expressed as $e_i = g\left(\mathcal{F}\left(x_i; \theta\right)\right)$. Here, $\mathcal{F}$ is the white-box model parameterized by θ and g is the evidence function such as exp, softplus, or ReLU.

Finally, the evidence of non-target classes is suppressed by minimizing the KL divergence between the modified Dirichlet distribution and the uniform distribution. Specifically, the regularization term has the following form:

$$\mathcal{L}_{kl} = \mathrm{KL}\left(\mathrm{Dir}\left(\mathbf{p}, \tilde{\alpha}_{x_i}\right), \mathrm{Dir}\left(\mathbf{p}, \mathbf{1}\right)\right) , \tag{B.6}$$

where $\mathrm{Dir}\left(\mathbf{p}, \mathbf{1}\right)$ is the uniform Dirichlet distribution, $\tilde{\alpha}_{x_i} = y + (1-y) \odot \alpha_{x_i}$ is the Dirichlet parameter for sample x_i after removing non-misleading evidence from the predicted parameters, and $\odot$ represents the Hadamard product. Therefore, the overall loss function is as follows:

$$\begin{aligned}\mathcal{L}\left(\theta\right) &= \textstyle\sum_{k=1}^{K} y_{ik}\left(\log\left(\mathcal{S}_i\right) - \log\left(e_{ik}+1\right)\right) \\ &\quad + \lambda_t \sum_{i=1}^{m} KL\left[\mathrm{Dir}\left(\mathbf{p_i} \mid \tilde{\alpha}_{x_i}\right) \parallel \mathrm{Dir}\left(\mathbf{p_i} \mid \langle 1, \cdots, 1\rangle\right)\right] ,\end{aligned} \tag{B.7}$$

where $\lambda_t = \min\left(1.0, t/10\right) \in [0, 1]$ is the annealing coefficient, and t is the index of the current training epoch.

[7] In the fine-tuning instruction set, we set the answer as a binary classification problem (yes or no), considering the top-5 most likely tokens generated in the answer, where $K = 5$.

[8] Please note that, similar to existing research [4], we choose the NLL loss as the loss function. However, in other scenarios, the determination of $\mathcal{L}_{edl}$ should be flexibly chosen based on the specific task to achieve optimal model performance.

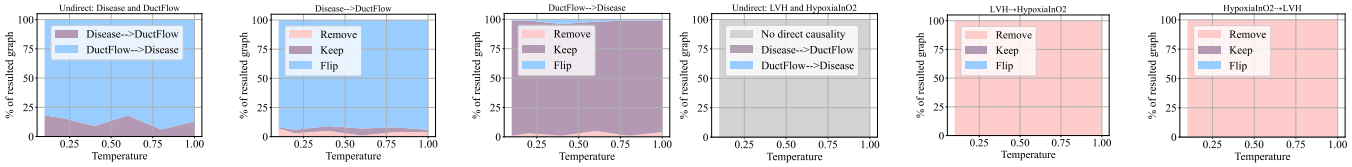


Figure 8: Hypothesis amendments across two different types of relationships: relationships that plausibly exist in a single direction (such as Disease and DuctFlow) and variables that have no biologically plausible causal link (LVH and HypoxiaInO2).

B.2 Derivation for Eq. 12

Combining the generative paradigm of LLMs, LLMs predicts the next token from the training corpus based on the preceding tokens. The output is a vector of the size of the corpus, representing the probability of each word unit becoming the next token. We formatted the fine-tuning data into an instruction set format and fine-tuned the white-box model to enhance its causal reasoning capabilities. Note that, due to time and cost constraints, we did not attempt to fine-tune larger models with more parameters, which will be part of our future work. We treat each query in our instruction set as a classification task, and its corresponding uncertainty can be calculated using the Equation 12. Below, we will derive the expression step by step. The output vector of the second-to-last layer of the white-box LLMs is regarded as the evidence vector in the theory of EDL. For a query x_i in the instruction set, the corresponding evidence vector $\mathbf{e}$ is $\sigma(f(x_i))$.

$$u_i = \frac{K_i}{\mathcal{S}_i} = \frac{K_i}{\sum_{l=1}^{K_i} \alpha_l} = \frac{K_i}{\sum_{l=1}^{K_i} \mathbf{e}_l + 1} = \frac{K_i}{\sum_{l=1}^{K_i} \sigma(f(x_i))_l + 1}, \tag{B.8}$$

where u_i represents the uncertainty corresponding to the generated i-th answer.

C MORE EXPERIMENTAL RESULTS

C.1 Human Evaluations

The specific evaluation process is as follows: 1. Medical experts score each edge in the DAG on a scale of 1 to 10, with higher scores indicating a higher likelihood of correctness. 2. Provide the reasoning and evidence for the rating, including Randomized Controlled Trials (RCT), Cohort Study, Mendelian Randomization Analysis, and Clinical Guidelines Recommendation. 3. Perform statistical analysis of the evaluation results.

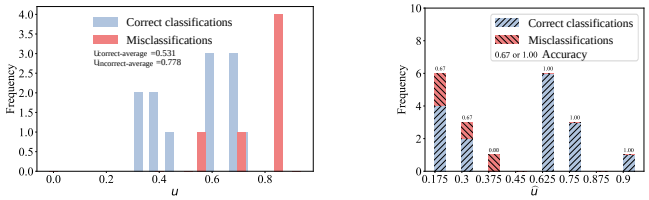
C.2 Behavioural experiment

The second half of Figure 8 illustrates a relationship that is unlikely to exist in either direction: LVH and HypoxiaInO2. Indeed, we observe that if this edge does not exist, LLM-CD does not introduce it; if the edge already exists (e.g., from a previous iteration), it is almost always removed. Therefore, the LLM module of LLM-CD exhibits a broadly expected behavior: when relationships are not fully understood, it typically defaults to the current input hypotheses. However, LLM-CD can also make suggestions that may not necessarily align with expert consensus. Relationships that should not exist at all are not suggested or are removed.

C.3 Uncertainty Estimation

To conduct uncertainty estimation for white-box LLMs such as LLaMA, we innovatively introduce the theory of EDL. We select the CORR2CAUSE (nodes $N = 4$) [17] dataset to construct the instruction set for fine-tuning. It should be noted that the goal of this paper is to deeply integrate LLMs and TCDA for causal discovery, rather than to fine-tune LLMs directly for causal discovery. Therefore, we only used a small-scale dataset and a relatively smaller model, LLaMA-3.1-8B, for fine-tuning. Fine tuning using large-scale datasets will be our future research work.

As shown in Figures 9, we further compare our uncertainty estimation method with the latest LLMs uncertainty methods [42, 55]. It can be seen that under the white-box LLM (LLAMA-3.1-8B), our uncertainty estimation method can assign lower uncertainty values to correctly classified samples, which [42] cannot achieve. Similarly, under the black-box LLM (GPT series), [55] cannot provide more accurate answers at lower uncertainty thresholds.



(a) [42] for LLAMA-3.1-8B **(b) [55] for GPT-4O**

Figure 9: The uncertainty distribution across different baseline on Child datasets.

C.4 Parameter sensitivity

Figure 10 shows the impact of the uncertainty threshold on model performance on the Child dataset. In LLAMA-3.1-8B, the model performance decreases as the uncertainty threshold increases. In GPT-4O, the model performance increases as the threshold increases.

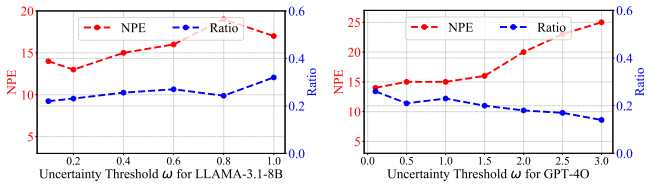


Figure 10: Performance with different uncertainty threshold ω on the Child dataset.