LLM-Driven Causal Discovery via Harmonized Prior

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, and Huanhuan Chen, *Fellow, IEEE*

***Abstract*—Traditional domain-specific causal discovery relies on expert knowledge to guide the data-based structure learning process, thereby improving the reliability of recovered causality. Recent studies have shown promise in using the Large Language Model (LLM) as causal experts to construct autonomous expert-guided causal discovery systems through causal reasoning between pairwise variables. However, their performance is hampered by inaccuracies in aligning LLM-derived causal knowledge with the actual causal structure. To address this issue, this paper proposes a novel LLM-driven causal discovery framework that limits LLM's prior within a reliable range. Instead of pairwise causal reasoning that requires both precise and comprehensive output results, the LLM is directed to focus on each single aspect separately. By combining these distinct causal insights, a unified set of structural constraints is created, termed a harmonized prior, which draws on their respective strengths to ensure prior accuracy. On this basis, we introduce plug-and-play integrations of the harmonized prior into mainstream categories of structure learning methods, thereby enhancing their applicability in practical scenarios. Evaluations on real-world data demonstrate the effectiveness of our approach.**

***Index Terms*—Causal discovery, expert-based system, large language model.**

I. Introduction

Structure learning of directed acyclic graph (DAG) models from observational data is an essential process for causal discovery in scientific contexts [1], [2]. However, the quality issue of real-world data and the NP-hard nature of the structure learning problem challenge the output reliability by data-based algorithms [3], [4]. Consequently, experts are heavily relied upon to help recover genuine causal mechanisms in practice, such as providing structural [5] and parametric constraints [6]. This dependency on expert labor leads to a barrier between structure learning techniques and scientific application fields, which greatly limits its practical utility [5].

A study by Kıcıman et al. [7] observes a promising capability of the Large Language Model (LLM) in knowledge-based causal reasoning tasks. Some studies explore using this causal reasoning ability to enhance data-based causal analysis, showing a promising frontier of integrating LLM in causal discovery [8], [9], [10]. On this basis, many studies further propose different ideas of LLM-driven causal discovery approaches and advance this line of research [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. These methods use LLM to reason about causal relationships between variable pairs and specify structural constraints accordingly, providing a flexible and autonomous way to inform data-based structure learning with causal knowledge. This approach frees the need for interdisciplinary experts with proficiency in both structure learning and the scientific domain, thereby breaking down barriers to using structure learning for reliable causal analysis.

However, LLM-driven structure learning carries inherent risks of prior inaccuracies. For example, some methods take the LLM-reasoned causal relationship between the variable pair to specify presence of edges [12], [18], [20]. Since the edges of a causal DAG are usually sparse [21] and the LLM is less reliable than human experts, this approach can easily introduce erroneous edges, compromising the recovered structure. To alleviate this risk, some studies use looser[1] structural constraints, such as ancestral and order constraints[2] [9], [11], which yet do not fundamentally address inaccurate prior constraints.

These inaccuracies of LLM-derive priors essentially arise from the misalignment between knowledge-reasoned causality and actual causal mechanisms underlying data. First, causality derived from knowledge is qualitative and can be indirect, which does not align with the edges of the graphical model that represent direct interactions. Additionally, some variables can be connected by knowledgeably unclear mechanisms, making the knowledge-reasoned results unreliable, even within the scope of indirect causality. For common strategies that rely on LLMs to infer causality between all pairs of variables [7], [12], [16], and then directly use these inferred relationships to define edge existence, there is a significant risk of introducing erroneous edges. Such errors can negatively impact the accuracy and reliability of the recovered causal structure.

To address the inaccuracy of knowledge-reasoned causality in revealing actual structures, this paper proposes a novel LLM-driven causal discovery framework, which limits the LLM's influence on structure learning to the range of reliable knowledge. Instead of providing specific answers on the causal relationships of arbitrary variable pairs, which requires both precise and comprehensive causal knowledge among the investigated variables, the LLM is directed to focus on each aspect separately. Through

Received 1 July 2024; revised 6 November 2024; accepted 7 January 2025. Date of publication 13 January 2025; date of current version 7 March 2025. This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0111700, in part by the National Nature Science Foundation of China under Grant 62137002, Grant 62176245 and Grant 62406302, and in part by the Natural Science Foundation of Anhui province under Grant 2408085QF195. Recommended for acceptance by X. Zhu. *(Corresponding authors: Xiangyu Wang; Huanhuan Chen.)*

The authors are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: sa312@ustc.edu.cn; hchen@ustc.edu.cn).

Digital Object Identifier 10.1109/TKDE.2025.3528461

[1]Consider structural constraints as binary relations. We say that a binary relation B_1 is looser than another B_2 if $B_2(u, v)$ implies $B_1(u, v)$.

[2]Ancestral constraints indicate that a directed path exists between two variables, whereas an order constraint specifies that one variable precedes another in the topological ordering of nodes.

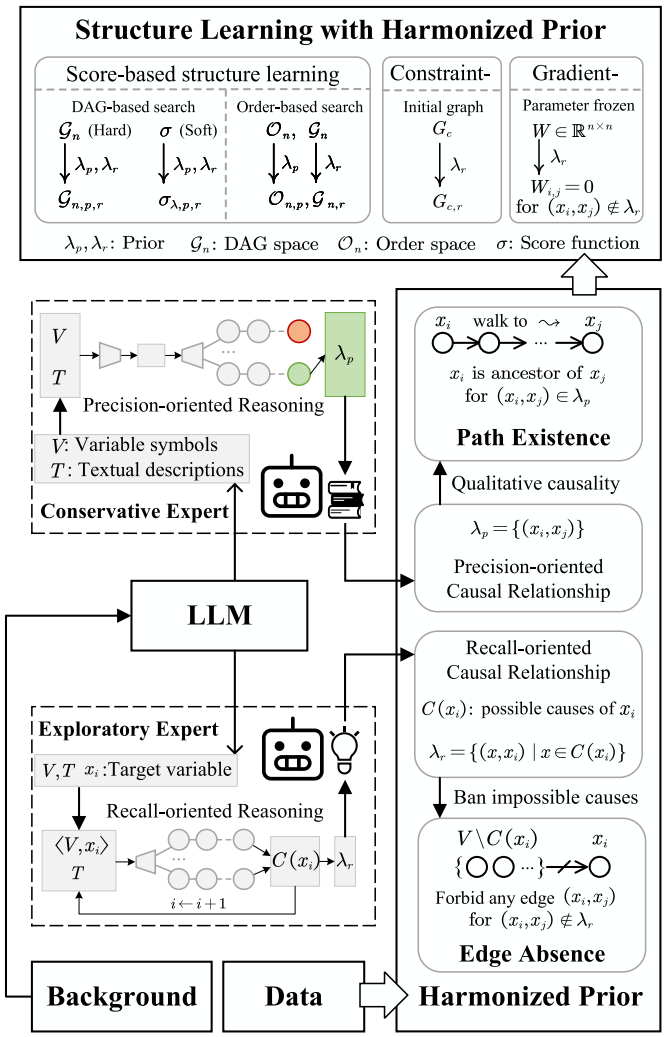


Fig. 1. Overview of the LLM-driven structure learning framework. The thin arrow represents the data flow of the LLM-based prior derivation process, while the thick arrow denotes the prior constraint-based structure learning process. The block of 'Background' illustrates variable descriptions and domains. This semantic information is input into the LLM performing two distinct roles of causal experts to derive a harmonized prior on causal relationships. Blocks with dotted line represents LLM's intermediate reasoning steps, which are individually designed to generate different aspects of causal information to form the harmonized prior. This harmonized prior is then used as structural constraints in the structure learning process. We investigate the integration of this prior across mainstream lines of structure learning methods, as shown in the top block.

specialized prompting strategies, the LLM can derive two distinct aspects of causal knowledge with better alignment to the actual structure. These two causal insights are then unified into a *harmonized* prior to inform structure learning. Subsequently, we delve into illustrations of the process of prompting and structure learning.

In the prompting process, we configure the LLM to individually act as a conservative and an exploratory expert to derive precision- and recall-oriented priors, respectively. The conservative expert is initially tasked with extracting obvious causal relationships through single-step reasoning, followed by decomposition into individual verifications[3] to filter out errors, thus ensuring precision. The exploratory expert focuses on identifying any potential influence on each variable, providing a comprehensive set of possible causal links.

In the structure learning process, we construct a harmonized constraint set that takes the precision-oriented prior as indirect causality while restricting hypotheses as subsets of the recall-oriented prior. On this basis, we design plug-and-play strategies to apply this harmonized prior in mainstream categories of structure learning methods, including score-based, constraint-based, and gradient-based approaches. These methods encompass the main causal graphical models, Bayesian networks [24], [25], and structural equation models [26], [27], providing comprehensive utility in practical scenarios.

An overview of the proposed LLM-driven structure learning framework is presented in Fig. 1. Evaluations on six real-world structures demonstrate significant improvements of our LLM-driven structure learning over various algorithms. The contributions of this paper are concluded as follows:

- This paper proposes a novel approach to integrate LLM experts in causal discovery, enhancing the reliability of LLM-derived priors by separating causal reasoning into respective precision and recall aspects.
- This paper illustrates converting the precision- and recall-oriented causal insights into harmonized prior constraints, which captures their respective strengths in knowledge aspects of causality to ensure the prior accuracy.
- This paper provides plug-and-play approaches to integrate the harmonized prior into mainstream categories of structure learning algorithms, which ensures wide applicability of our method in practical scenarios.

The rest of the contents is organized as follows. Section II systematically reviews the emerging research on LLM-driven causal discovery. Section III presents preliminaries. Section IV introduces our framework in the context of diverse lines of structure learning. Section V reports experimental results and conducts analysis. Section VI makes conclusions.

II. Related Work

This section introduces the line of most related studies with the topic of LLM and causal discovery. These studies are broadly split into two stages, one for LLM's inherent ability in causality and one for its integration in data-based structure learning for causal discovery.

A. LLM's Causal Reasoning Ability

Reasoning ability is a critical aspect of foundation models [28]. Therefore, it is intuitive to leverage this reasoning ability in causal tasks, as explored in numerous studies.[4] Zhang et al. [29] explored the ability of LLM to answer questions of causality and found that while they excel at answering questions

[3] Autonomously decomposing a complex task into smaller ones improves reasoning reliability [22], [23].

[4] Note that the majority of studies introduced here use GPT-3 or GPT-3.5-turbo for their statements, whose performance can be different from current state-of-the-art LLMs like GPT-4, Llama-3, etc.

based on existing causal knowledge, they fall short in discovering new knowledge. Kıcıman et al. [7] reviewed a collection of research on LLM's capabilities in various causal tasks [30] and conducted further evaluations, observing promising performance in causal analysis between pairwise variables. Naik et al. [31] noted the good capability of LLM in identifying causal directionality in a medical dataset. These results showcase promise of LLM's reasoning ability in identifying causal relationships in certain domains.

However, the LLM is not as effective in constructing causal graphs. Long et al. [32] demonstrated that causal medical graphs constructed solely using LLM are not sufficient to replace expert judgment. Tu et al. [30] found that the LLM struggles to provide accurate causal graphs among new concepts in medicine. Additionally, Zečvi'c et al. [33] argued that the LLM merely preserves causal knowledge from their training texts rather than truly processing causal reasoning. Romanou et al. [34] tested several LLMs on a new causal reasoning assessment benchmark and observed that they struggle with recognizing untrained causal knowledge. Similar observations have been verified in related studies [35], [36]. These results show that the LLM is limited to the range of existing causal knowledge and struggle to identify new causal knowledge outside of their training data. Therefore, the LLM cannot solely handle causal discovery without data-based analysis.

B. Integration of LLM in Causal Discovery

Beyond solely using LLM to reason about knowledge-based causality, some studies have begun treating this information as priors in data-based causal analysis. Choi et al. [8] introduced the use of LLM-based priors in several data-driven tasks and observed promising enhancements, including inferring the causal directionality of pairwise variables. Another study by Long et al. [10] explored using LLM-based priors to identify causal directions within an assumed correct structural skeleton, which is helpful as a post-hoc process to identify a unique output from the equivalent DAGs generated by structure learning. While these studies do not directly focus on the causal discovery task, they demonstrate the potential of LLM in enhancing data-based causal analysis.

As an initial exploration of integrating LLM directly in causal discovery, Ban et al. [9] employed the LLM-derived causal relationships as path existence in two structure learning algorithms, with both improved and degraded performances observed across various datasets. To mitigate inaccuracies of LLM-based priors, Vashishtha et al. [11] introduced a voting process to filter out suspicious priors. Abdulaal et al. [12] used an iterative strategy to refine structure learning with LLM by verifying each pair of variables. Li et al. [16] adapted LLM-driven structure learning to temporal cases, proposing a prior structure-based temporal causal discovery method. Further studies on LLM-driven structure learning for causal discovery have been conducted along similar lines [14], [18], [19], [20]. For more details, one can refer to related surveys [28], [37], [38], [39].

In these studies, a critical concern exists regarding inaccuracies in the consistency between knowledge-based causal reasoning and data-implied causal structures. Some error-resistant mechanisms have been explored to alleviate this issue, yet they fail to address the essential limitation of knowledge-based causality. Therefore, this paper focuses on this critical aspect to enhance LLM-driven causal discovery.

III. Preliminaries

This section broadly introduces the main structure learning methods, serving as a preliminary of the backbone algorithms used in our LLM-driven structure learning framework.

A. Causal Graphical Model

Bayesian networks (BNs) [40], [41], [42], [43], [44], [45] and structural equation models (SEMs) [46] are two most popular DAG models to represent causality. Suppose a DAG $G(V, E)$, where $V = \{x_i\}_{i=1}^n$ is node set corresponding to random variables, and $E \subset V \times V$ is the edge set. A BN models the data distribution markovly:

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p\left(x_i \mid \mathrm{Pa}^G\left(x_i\right)\right) \tag{1}$$

where $\mathrm{Pa}^G(x_i) = \{y \mid (y, x) \in E(G)\}$ is the set of parent nodes of x_i in the DAG.

An SEM models the causality as a set of functions:

$$x_i = f_i\left(\mathrm{Pa}^G\left(x_i\right), z_i\right) \tag{2}$$

where z_i is a noise term. Given a set of samples $D \in \mathbb{R}^{m \times n}$ generated by the graphical model, structure learning is to recover the graph structure from data D. Broadly speaking, we use BNs for discrete data and SEMs for continuous data. Subsequently, we introduce main structure learning methods.

B. Overall Review on Main Structure Learning Methods

Mainstream structure learning approaches include score-based [40], constraint-based [21], and hybrid [47] methods typically used for BNs, and gradient-based [46] and Independent Component Analysis (ICA)-based [48] methods typically used for SEMs. Notably, ICA-based methods are specifically designed for linear SEMs with non-Gaussian noise [48], making them unsuitable for general cases, and thus they are not included in the scope of this paper. Hybrid methods are a combination of score-based and constraint-based approaches [47]. Therefore, this section mainly introduces score-based, constraint-based, and gradient-based methods.

Score-based Methods: This method evaluates a graph's goodness-of-fit to the data and searches for the optimal DAG [40], [49]. Well-known scores include the Bayesian Information Criterion (BIC) [50], Bayesian Dirichlet equivalent uniform (BDeu) [51], and Minimum Description Length (MDL) [52]. These scores are typically decomposable into local scores due to the Markov property of BNs:

$$\sigma(G; D) = \sum_i \sigma\left(x_i, \mathrm{Pa}_i^G; D\right) \tag{3}$$

With such scores, the following constrained combinatorial optimization problem is addressed to derive the solution:

$$\min_{G} \sum_{i} \sigma\left(x_i, \mathrm{Pa}_i^G; D\right) \quad \text{s.t. } G \in \text{DAG} \tag{4}$$

Representative methods include greedy equivalence search (GES) [53], hill climbing (HC) [54], Tabu [55], and A* [56], etc.

In addition to direct search in the DAG space, order-based search is also a significant strategy to solve this problem [57]. It assumes the total ordering π of variables to ensure adherence to the acyclicity constraint. By this means, the problem in (4) is reduced into n unconstrained local searches:

$$\min_{\mathrm{Pa}_i^G \subseteq \pi_i} \sigma\left(x_i, \mathrm{Pa}_i^G; D\right) \quad i = 1, 2, \ldots, n \tag{5}$$

where π_i are the set of variables that precede x_i. These local search problems can be efficiently addressed to derive the optimal solution of an ordering. On this basis, the order space is searched to find the optimal DAG globally.

Constraint-based Methods: This method constructs the graph by conditional independence (CI) tests. The most representative algorithm, the PC algorithm [21], starts from a complete graph and removes edges whose nodes are dependent given certain variables. This algorithm outputs partial DAGs, whose skeleton is further oriented in a post-hoc process. Other algorithms include fast causal inference (FCI) [58] and some variant PC algorithms [59].

Gradient-based methods: This method differs significantly from the aforementioned methods. It converts structure learning from a combinatorial optimization problem to a continuous one by a continuous characterization of the acyclicity constraint [46]. Thus, structure learning is formalized as an equality-constrained problem (ECP):

$$\min_{W \in \mathbb{R}^{n \times n}} F(W; D) \quad \text{subject to } h(W) = 0$$
$$h(W) = \text{Trace}\left(e^{W \circ W}\right) - d \tag{6}$$

Here, W is the weighted adjacency matrix of the graph, where $W_{i,j} \neq 0 \iff (x_i, x_j) \in E(G)$. $F(W; D)$ is the goodness-of-fit score of the graph to the data. $h(W) = 0$ represents the acyclicity constraint. The definition of $h(W)$ varies [60], [61], and (6) is the first proposed one [46]. This ECP can be addressed using various numerical optimization methods, with the augmented Lagrangian method [62] being the most popular in gradient-based approaches. Other representative algorithms include DAG-GNN [60], DAGMA [61], etc.

IV. Structure Learning With LLM-Derived Harmonized Prior

In this section, we present our proposed LLM-driven structure learning framework. We begin by detailing the process of deriving the harmonized prior from the LLM. Subsequently, we explain how this prior can be utilized as structural constraints in various structure learning algorithms. To set the stage, assume the variable set is $V = \{x_i\}_{i=1}^n$. A graph with these n variables as nodes is denoted as $G(V, E(G))$, where $E(G) \subseteq V \times V$ is the edge set of G.

A. Derivation of Harmonized Prior From LLM

This section introduces how to derive the harmonized prior on causality from LLM by specialized prompting. Given a variable set $V = \{x_i\}_{i=1}^n$, their descriptive texts $T = \{t_i\}_{i=1}^n$, and background information, the prior extracted by LLM is a set of pairwise causality $\{(u, v) \mid u, v \in V\}$.

To derive reliable causal reasoning results, we make the LLM to focus separately on the precision and recall aspects of causal reasoning, termed conservative ($\mathcal{A}$) and exploratory ($\mathcal{E}$) modes, respectively. $\mathcal{A}$ prioritizes the accuracy of the extracted prior, while $\mathcal{E}$ identifies a comprehensive set of any possible causality. These priors do not aim to exactly identify every causality in the actual structure, some of which can be knowledgeably unclear. Instead, they are reduced to the range of clear causality, thus ensuring the reliability of reasoning. This clear causality is derived in both positive (precision) and negative (recall) aspects to maximize the prior quantity.

To derive the precision- and recall-oriented priors, we design respective prompting strategies to control the LLM within these scopes of causal reasoning. The common setup includes providing background information on the variables and their respective behaviors, specifying input and output formats, and tailoring requirements for extracted causality. Next, we introduce the distinct parts of the two prompting strategies.

Conservative Mode: The LLM in the conservative mode is prompted with a focused single-step and decomposed verification strategy to prioritize the precision. It is first fed with all the variables V and descriptions T to conduct a single-step reasoning on causality extraction. This operation aids in focusing on obvious causal relationships (x_i, x_j) as output instead of delving into analysis of pairwise variables that may produce fake causality. This aligns with our expectation on the assurance of correctness of the causality. Moreover, compared to pairwise analysis, this operation reduces the LLM reasoning complexity from $O(n^2)$ to $O(1)$.

The causal relationships derived by the single-step reasoning is subsequently decomposed into piece-to-piece verification with chain-of-thought (CoT) [63] reasoning. This process is designed to handle the lost of reliability of LLM outputs when handling complex problems by directly providing answers, as evidenced in related studies [22], [63]. Additionally, a re-confirmation mechanism is designed to check the reasoning-level correctness, specifically the reasoning trace of each piece of causality verification. It aims to address the distortion of the reasoning traces of the LLM, where certain intermediate step of the LLM reasoning process can be distorted by incident thus harming the result, as evidenced in relevant research [23]. The causal relationships re-confirmed to be true is taken as ouput of the conservative LLM expert.

By prompting the LLM with this strategy, we derive the precision-oriented prior, denoted as:

$$\lambda_p = \mathcal{A}(V, T) = \{(u, v) \mid u, v \in V \text{ and}$$

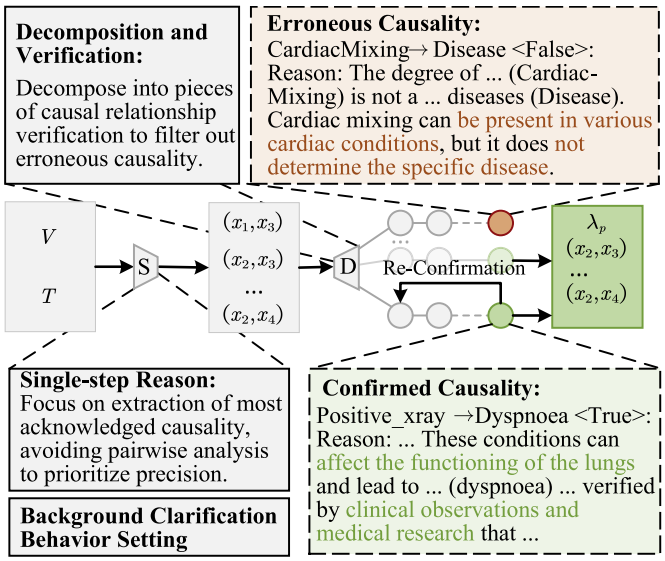


Fig. 2. The causal reasoning process of the conservative LLM expert. Zoom-in boxes with solid borderlines illustrate the controlled behavior of LLM reasoning, those with dotted borderlines are examples of LLM output. Circles represent the the reasoning trace of LLM. Symbols can be referred to in Section IV-A.

$$(u, v) \text{ is true as causality by } \mathcal{A}\} \tag{7}$$

An illustrative diagram and example are available Fig. 2.

Exploratory Mode: The LLM in the exploratory mode is prompted with an incremental potential cause identification strategy to prioritize recall. We decompose the complex task of identifying causal relationships among variables V into n simpler sub-tasks to identify possible causes of individual variables x_i from V. As illustrated before, this step makes LLM to conduct more fine-grained and reliable reasoning. Subsequently, we further make LLM decompose into analyzing potential causal relationships between each variable in V and the target x_i. Different from the conservative LLM expert, we do not make a re-confirmation operation here to avoid mis-removing true causal relationships, as our expectation here is to cover all true causality as possibly.

All the variables identified to be possibly influencing target variable x_i are taken as its possible cause set $C(x_i)$. By n rounds of prompting possible causes of each variable, these causal relationships are incrementally identified and taken as the recall-oriented prior:

$$\lambda_r = \bigcup_{i=1}^{n} \{(u, x_i) \mid u \in C(x_i)\}$$
$$C(x_i) = \mathcal{E}(V, x_i, T) = \{u \mid u \in V \text{ and } u \text{ is true as a possible cause of } x_i \text{ by } \mathcal{E}\} \tag{8}$$

An illustrative diagram and example are available Fig. 3.

By these prompting strategies, LLM-reasoned causality in λ_p and λ_r prioritize the precision and recall with respect to the actual causal structure, respectively. Leveraging their respective

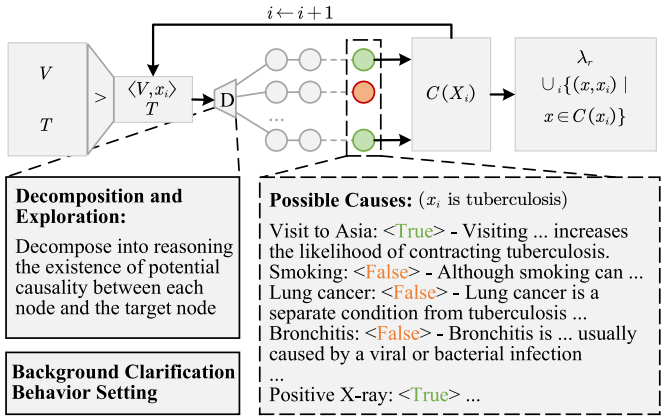


Fig. 3. The causal reasoning process of the exploratory LLM expert.

strengths, we unify them into a harmonized prior λ_H that corresponds to the following graph space:

$$\mathcal{G}_{n,H} = \{G \mid \forall (x_i, x_j) \in \lambda_p \cap \lambda_r,\ x_i \rightsquigarrow x_j \text{ in } G \text{ and } E(G) \subseteq \lambda_r \cup \lambda_p\} \tag{9}$$

where n is the number of nodes, $x_i \rightsquigarrow x_j$ indicates a directed path from x_i to x_j, and $E(G)$ is the edge set of G.

The harmonized prior λ_H employs the precision-oriented prior λ_p to specify the existence of paths (qualitative causality) and the recall-oriented prior λ_r to restrict the range of edges (direct causality). The set operation of λ_p and λ_r in (9) addresses contradictory reasoned causality, such as when a causality is confidently identified in λ_p but not included in the possible causality in λ_r, and vice versa. These cases indicate an unclear recognition of knowledge; therefore, we remove such unreliable priors from λ_H. Additionally, λ_p should be acyclic, where all pairs that form a cycle are removed.

Notably, we refer to the path existence constraints specified by λ_p and edge absence constraints specified by λ_r in the following illustrations. Moreover, restricting the edges of G within $\lambda_r \cup \lambda_p$ is equivalent to forbidding edges out of it, both of which represent the constraints of λ_r. An example of prompts is presented in Table I.

B. Plug-and-Play Application of Harmonized Prior in Score-Based Methods With Hard and Soft Constraints

This section introduces how to integrate the LLM-derived harmonized prior into arbitrary score-based methods defined by a decomposable scoring function σ and search method $\mathcal{M}$. We introduce two integration strategies, a hard approach that strictly ensures the output is in the constrained graph space of (9) and a soft approach that balances the adherence to priors and the goodness-of-fit to data.

Given data $D \in \mathbb{R}^{m \times n}$, the process by which the score-based method derives the solution can be expressed as follows:

$$\min_{G \in \mathcal{G}_n} \sigma(G, D) \quad \text{s.t. } G \in \text{DAG} \tag{10}$$

$$G^{\text{opt}} \leftarrow \mathcal{M}_{\mathcal{G}_n}(D, \mathcal{N}, \sigma) \quad \text{solves (10)} \tag{11}$$

TABLE I
AN EXAMPLE OF THE PROMPTS USED TO CONSTRUCT CONSERVATIVE AND EXPLORATORY LLM EXPERTS FOR CAUSAL REASONING

Component	Description
Reasoning of Precision-Oriented Causality (Conservative LLM Expert)	
Single-Step Causal Reasoning	You are an expert in `[Background Clarification]`. Based on your expertise, please extract the causal relationships among the following variables that you are confident in: $x_1 : t_1, \cdots, x_n : t_n$. Output `[Behaviour Setting]`.
Decomposition and Verification	Please verify the correctness of the extracted causal relationships step by step. Output `[Behaviour Setting]`.
Re-Confirmation	`For each verified causal relationship` (x_i, x_j)`:` Analyze the causal relationship between x_i and x_j. Output `[Behaviour Setting]`.
Recall-Oriented Causality (Exploratory LLM Expert)	
Decomposition and Exploration	`For each variable` x_i `in the set of variables:` You are an expert in `[Background Clarification]`. Identify the potential causes for the variable $x_i : t_i$ from the following: $x_1 : t_1, \cdots, x_n : t_n$. Output `[Behaviour Setting]`.

Algorithm 1: Hard Score-Based Structure Learning.

Require: Data $D \in \mathbb{R}^{m \times n}$, Harmonized Prior $\lambda_H = (\lambda_p, \lambda_r)$, Scoring Function σ, Search Method $\mathcal{M}$, Graph Neighbor Definition $\mathcal{N}$
Ensure: Learned Structure G^{opt}
1: Define Revised Graph Neighbor Definition $\mathcal{N}_r(G)$ as follows:
2: **function** RevisedNeighbors $\mathcal{N}_r G$
3: **return** $\{G' \mid E(G') = E(G) \cup \{(u,v)\} \setminus \{(p,q)\}$ for some $(u,v) \in \lambda_r \cup \lambda_p \setminus E(G), (p,q) \in E(G)\}$
4: **end Function**
5: Define Revised Scoring Function $\sigma_p(G, D)$ as follows:
6: **function** RevisedScoring $\sigma_p G, D$
7: Calculate $\sigma(G, D)$ using the original scoring function.
8: Calculate $\mathcal{P}_{\lambda_p}(G) = \sum_{(x_i,x_j)\in\lambda_p\cap\lambda_r} \mathbf{1}(\nexists x_i \rightsquigarrow x_j \text{ in } G)$
9: **return** $\sigma(G, D) + \inf \times \mathcal{P}_{\lambda_p}(G)$
10: **end Function**
11: $G^{\text{opt}} \leftarrow$ Run Search Method $\mathcal{M}(D, \mathcal{N}_r, \sigma_p)$
12: **return** G^{opt}

where σ is the scoring function, $\mathcal{G}_n$ is the space of graphs with n nodes, $\mathcal{N}$ is the definition of graph neighbors. Note that the subscript of $\mathcal{M}$ is merely a notion indicating its hypothesis space, used to highlight the impact of priors on search. Next, we detail the hard and soft approaches to apply λ_H in $\mathcal{M}$.

Hard Approach: The hard approach to apply the harmonized prior aims to restrict the graph space $\mathcal{G}_n$ to $\mathcal{G}_{n,H}$ as defined in (9). To restrict this graph space, we delve into the definition of graph neighbors $\mathcal{N}$:

$$\mathcal{N}(G) = \{G' \mid E(G') = E(G) \cup \{(u,v)\} \setminus \{(p,q)\} \text{ for some } (u,v) \notin E(G),\ (p,q) \in E(G)\} \quad (12)$$

The neighbor set of a graph is generated by removing, adding, or reversing some edges. The constraints of λ_r can be integrated in this step by revising the neighbor definition:

$$\mathcal{N}_r(G) = \{G' \mid E(G') = E(G) \cup \{(u,v)\} \setminus \{(p,q)\} \text{ for some } (u,v) \in \lambda_r \cup \lambda_p \setminus E(G),\ (p,q) \in E(G)\} \quad (13)$$

This revision ensures that all searched graphs[5] maintain an edge set $E(G) \subseteq \lambda_r \cup \lambda_p$.

For the path existence constraints specified by λ_p, they are applied by revising the scoring function σ:

$$\sigma_p(G, D) = \sigma(G, D) + \inf \times \mathcal{P}_{\lambda_p}(G)$$
$$\mathcal{P}_{\lambda_p}(G) = \sum_{(x_i,x_j)\in\lambda_p\cap\lambda_r} \mathbf{1}(\nexists x_i \rightsquigarrow x_j \text{ in } G) \quad (14)$$

Here, $\mathbf{1}$ is the indicator function that equals 1 if the condition inside is true and 0 otherwise. $\mathcal{P}_{\lambda_p}(G)$ denotes the number of path existence constraints not satisfied by G, and $\inf$ is a sufficiently large value to ensure the sub-optima of prior-violating graphs compared to the satisfying ones. This revision ensures that the path $x_i \rightsquigarrow x_j$ is present in the optimal DAG for all (x_i, x_j) in $\lambda_p \cap \lambda_r$ as long as feasible DAGs exist.

As a result, we derive the hard plug-and-play application of the LLM-derived harmonized prior in arbitrary score-based methods, defined by scoring function σ, search method $\mathcal{M}$, and graph neighbor definition $\mathcal{N}$, formalized as follows:

$$\mathcal{M}_{\mathcal{G}_n}(D, \mathcal{N}, \sigma) \xrightarrow{\lambda_H} \mathcal{M}_{\mathcal{G}_{n,H}}(D, \mathcal{N}_r, \sigma_p) \quad (15)$$

The revised definition of graph neighbors $\mathcal{N}_r$ (13) ensures the path existence constraints, and the revised scoring function σ_p (14) ensures the edge absence constraints. These revisions restrict the graph space to $\mathcal{G}_{n,H}$, ensuring strict adherence to the harmonized prior. The algorithm of hard score-based structure learning is presented in Algorithm 1.

Soft Approach: The soft approach requires a prior probability $c \in (0, 1)$ for each structural constraint. It is positive (specifying the existence of a structure) when $c \in (0.5, 1)$, negative when $c \in (0, 0.5)$, and inoperative when $c = 0.5$. Here, we unify the probability of all constraints as a confidence level c in (0.5,1), where the probability of path existence in λ_p is c, and probability of existence of edges out of λ_r is $1 - c$. This probability is integrated in the scoring function to balance the adherence to priors and goodness-of-fit to data.

As the scoring function typically estimates the logarithm of the posterior probability of DAG models, we consider the

[5]The initial graph is also restricted that $E(G) \subseteq \lambda_r \cup \lambda_p$.

Algorithm 2: Soft Score-Based Structure Learning.

Require: Data $D \in \mathbb{R}^{m\times n}$, Harmonized Prior $\lambda_H = (\lambda_p, \lambda_r)$, Scoring Function σ, Search Method $\mathcal{M}$, Confidence Level $c \in (0.5, 1)$
Ensure: Learned Structure G^{opt}
1: Define Revised Scoring Function $\sigma_{\lambda,H}(G, D)$ as follows:
2: **function** RevisedScoring $\sigma_{\lambda,H} G, D$
3: Calculate $\sigma(G, D)$ using the original scoring function.
4: Calculate
$$b_{\lambda,p} = \sum_{(u,v)\in\lambda_p\cap\lambda_r} \log\left(c\mathbf{1}(u \rightsquigarrow v) + (1-c)\mathbf{1}(u \not\rightsquigarrow v)\right)$$
5: Calculate
$$b_{\lambda,r} = \sum_{(u,v)\in\bar{\lambda_r}\cap\bar{\lambda_p}} \log\left((1-c)\mathbf{1}(u \to v) + c\mathbf{1}(u \nrightarrow v)\right)$$
6: **return** $\sigma(G, D) + b_{\lambda,p} + b_{\lambda,r}$
7: **end Function**
8: $G^{\text{opt}} \leftarrow$ Run Search Method $\mathcal{M}(D, \mathcal{N}, \sigma_{\lambda,H})$
9: **return** G^{opt}

Algorithm 3: Order-Based Search With Harmonized Prior.

Require: Data $D \in \mathbb{R}^{m\times n}$, Harmonized Prior λ_p, Scoring Function σ, Order Search Method S_o, Graph Search Method S_g
Ensure: Learned Structure G^{opt}
1: Define Graph Search Method S_{g,π_i}:
2: **function** GraphSearch $S_{g,\pi_i}\pi, \sigma, D$
3: **for** each variable x_i in π **do**
4: Restrict Parent Node Set $\text{Pa}_i \leftarrow \{u \mid u \prec x_i \text{ in } \pi\}$
5: Apply λ_r: Reduce π_i to $\pi_{i,r} = \pi_i \cap C(x_i) \cup \{u \mid (u, x_i) \in \lambda_p\}$
6: **end for**
7: $G_\pi^{\text{opt}} \leftarrow$ Solve $\min_{\text{Pa}_i\subseteq\pi_{i,r}} \sigma(x_i, \text{Pa}_i, D)$ using S_g
8: **return** G_π^{opt}
9: **end Function**
10: Define Order Search Method $S_{o,\mathcal{O}_n}$:
11: **function** OrderSearch $S_{o,\mathcal{O}_n}\sigma, D$
12: Reduce Order Space $\mathcal{O}_n$ to $\mathcal{O}_{n,p}$ using λ_p: $\mathcal{O}_{n,p} = \{\pi \mid u \prec v \text{ in } \pi \text{ for } (u, v) \in \lambda_p \cap \lambda_r\}$
13: Run Graph Search S_{g,π_i} and order search S_o in $\mathcal{O}_{n,p}$ to obtain the optimal graph G_π^{opt} and order π^{opt}
14: **return** G_π^{opt}
15: **end Function**
16: $G^{\text{opt}} \leftarrow$ Run Order Search Method $S_{o,\mathcal{O}_n}$
17: **return** G^{opt}

prior-based logarithm probability:

$$\log p\left(G \mid D, \lambda\right) \sim \log p\left(D \mid G\right) + \log p\left(G \mid \lambda\right) \tag{16}$$

Here, $x \sim y$ means $x = y + a$ for constant a. The symbol λ denotes a certain structural constraint on G. Expressed in terms of the scoring function, the prior-based criteria can be formalized as follows:

$$\begin{aligned}\sigma_\lambda(G, D) &= \sigma(G, D) + b_\lambda \\ b_\lambda &= \log\left(c\mathbf{1}\left(G \models \lambda\right) + (1-c)\mathbf{1}\left(G \not\models \lambda\right)\right)\end{aligned} \tag{17}$$

Here, $G \models \lambda$ indicates that G satisfies λ, and c is the confidence level. The harmonized prior consists of constraints of path existence by λ_p and edge absence by λ_r, leading to the following revision on the scoring function:

$$\sigma_{\lambda,H}(G, D) = \sigma(G, D) + b_{\lambda,p} + b_{\lambda,r} \tag{18}$$

Here, $b_{\lambda,p}$ and $b_{\lambda,r}$ evaluate the adherence to constraints specified by λ_p and λ_r, respectively, calculated as follows:

$$\begin{aligned}b_{\lambda,p} &= \sum_{(u,v)\in\lambda_p\cap\lambda_r} \log\left(c\mathbf{1}\left(u \rightsquigarrow v\right) + (1-c)\mathbf{1}\left(u \not\rightsquigarrow v\right)\right) \\ b_{\lambda,r} &= \sum_{(u,v)\in\bar{\lambda_r}\cap\bar{\lambda_p}} \log\left((1-c)\mathbf{1}\left(u \to v\right) + c\mathbf{1}\left(u \nrightarrow v\right)\right)\end{aligned} \tag{19}$$

Here, $u \to v$ indicates that edge $(u, v) \in E(G)$, and $u \rightsquigarrow v$ means that a directed path from u to v exists in G. Conversely, $u \nrightarrow v$ and $u \not\rightsquigarrow v$ indicate the absence of an edge and a directed path, respectively. $\bar{A}$ is the complementary set of A.

Hence, we derive the soft approach to apply the LLM-derived harmonized prior in an arbitrary score-based method:

$$\mathcal{M}_{\mathcal{G}_n}(D, \mathcal{N}, \sigma) \xrightarrow{\lambda_H} \mathcal{M}_{\mathcal{G}^c_{n,H}}(D, \mathcal{N}, \sigma_{\lambda,p,r}) \tag{20}$$

In the revised scoring function $\sigma_{\lambda,H}$, adherence to more constraints in the harmonized prior improves the prior adherence terms $b_{\lambda,p}$ and $b_{\lambda,r}$, thus promoting satisfaction of these priors. Unlike the hard approach, which prunes all prior-conflicting hypotheses, the soft approach makes a trade-off between prior adherence and data fit. It adjusts the prior strength via the confidence level c, where a higher c increases the separation between prior-adherent and prior-conflicting points. Therefore, we use the symbol $\mathcal{G}^c_{n,H}$ to denote this prior-reshaped hypothesis space[6] with a confidence level[7] c. The algorithm of soft score-based structure learning is presented in Algorithm 2.

C. Application of Harmonized Prior in Order-Based Search

This section introduces a specialized application of λ_p in score-based methods where the search is conducted in the order space, as illustrated in Section III-B. Here, λ_p is used to reduce the order space, which is more efficient than applying the path existence constraint that requires additional calculations to check for path existence, with this order-based harmonized prior denoted as λ^o_H.

For each $(x_i, x_j) \in \lambda_p \cap \lambda_r$, we derive that x_i precedes x_j ($x_i \prec x_j$) in the order. Notably, both order and path existence constraints are transitive,[8] capturing identical prior information

[6]Here, no hypotheses are explicitly removed from the space, and the symbol $\mathcal{G}^c_{n,H}$ is used to describe the degree to which the new scoring function *splits* the prior-adherent points from the prior-conflicting ones.

[7]Note that $\mathcal{G}^1_{n,H}$ equals $\mathcal{G}_{n,H}$, reducing the soft approach to the hard one when the confidence level is set to 1.

[8]A binary relation B is transitive if $B(u, v)$ and $B(v, p)$ imply $B(u, p)$.

in terms of binary relations. The key difference is that the order constraint $x_i \prec x_j$ captures the relation in the order space instead of the graph space, which does not necessarily demand the existence of a directed path.

To begin with, we split an order-based search method into components of a scoring function σ, search method S_o conducted on the order space $\mathcal{O}_n$, search method S_g conducted on the graph space $\mathcal{G}_n$ within a total ordering $\pi \in \mathcal{O}_n$.

We first delve into S_g that solves the structure learning problem given a total ordering π:

$$\min_{\mathrm{Pa}_i \subseteq \pi_i} \sigma\left(x_i, \mathrm{Pa}_i, D\right) \tag{21}$$

Here, the DAG constraint is satisfied by restricting the parent node set Pa_i of x_i as a subset of $\pi_i = \{u \mid u \prec x_i \text{ in } \pi\}$, which is the set of nodes preceding x_i in π. Hence, we denote the process of solving this problem as follows:

$$G_\pi^{\text{opt}} \leftarrow S_{g,\pi_i}(\sigma, D) \quad \text{solves (21)} \tag{22}$$

In this process, we apply λ_r by reducing π_i to $\pi_{i,r} = \pi_i \cap C(x_i) \cup \{u \mid (u, x_i) \in \lambda_p\}$, where $C(x_i)$ is the set of possible causes of x_i in λ_r, as defined in (8).

Next, we consider the search process $\mathcal{S}_o$ in the order space, which searches for the ordering π such that $\mathcal{G}_\pi^{\text{opt}}$ has the optimal score, thus deriving the global optimal DAG:

$$\min_{\pi \in \mathcal{O}_n} \sigma\left(S_{g,\pi_i}(\sigma, D), D\right) \tag{23}$$

$$G^{\text{opt}} \leftarrow \mathcal{S}_{o,\mathcal{O}_n}\left(S_{g,\pi_i}(\sigma, D)\right) \quad \text{solves (23)} \tag{24}$$

In this process, we reduce the order space $\mathcal{O}_n$ to $\mathcal{O}_{n,p}$ to apply λ_p as order constraints:

$$\mathcal{O}_{n,p} = \{\pi \mid u \prec v \text{ in } \pi \text{ for } (u, v) \in \lambda_p \cap \lambda_r\} \tag{25}$$

Hence, we derive the application of the harmonized prior with specialized use of λ_p as order constraints to any order-based search methods defined by σ, $\mathcal{S}_o$, and $\mathcal{S}_g$:

$$\mathcal{S}_{o,\mathcal{O}_n}\left(S_{g,\pi_i}(\sigma, D)\right) \xrightarrow{\lambda_H^o} \mathcal{S}_{o,\mathcal{O}_{n,p}}\left(S_{g,\pi_{i,r}}(\sigma, D)\right) \tag{26}$$

By reducing the candidate parent node set π_i to $\pi_{i,r}$, edge absence constraints of λ_r are satisfied. By reducing the order space $\mathcal{O}_n$ to $\mathcal{O}_{n,p}$, the order constraints of λ_p are satisfied. Compared to using λ_p as path existence constraints in score-based methods ((15) and (20)), using order constraints does not modify the scoring function and avoids global check for path existence. This accelerates the search process by reducing both the order and graph space. Additionally, using order constraints introduces error tolerance to priors in λ_p. A path existence (u, v) is false if $u \not\rightsquigarrow v$, while an order constraint $u \prec v$ remains true if $u \not\rightsquigarrow v$ and $v \not\rightsquigarrow u$. The algorithm of order-based search with harmonized prior is presented in Algorithm 3.

Remark 1: Please note that the order-based search is fundamentally a score-based approach. The integration strategy described here follows a hard approach, but it is also feasible to apply a soft approach. Specifically, for an arbitrary graph G sampled from an ordering $\mathcal{O}$, we can define a prior-based score as follows:

$$\sigma_{\lambda,\mathcal{O}}(G; D) = \sigma(G; D) + b_{\lambda,r} + b_{\lambda,\mathcal{O}}$$

Algorithm 4: Constraint-Based Structure Learning.

Require: Data $D \in \mathbb{R}^{m \times n}$, Harmonized Prior λ_r, Conditional Independence Test Method $\mathcal{T}$, Search Strategy $\mathcal{M}$
Ensure: Learned Structure G^{opt}
1: Initialize Complete Undirected Graph G_c
2: Revise G_c to $G_{c,r}$: $E(G_{c,r}) = \{(u, v) \mid (u, v) \in \lambda_r \cup \lambda_p\}$
3: $G^{\text{opt}} \leftarrow$ Run Search Strategy $\mathcal{M}(G_{c,r}, D, \mathcal{T})$
4: **return** G^{opt}

where $\sigma(\cdot)$ represents the data approximation score, $b_{\lambda,r}$ is defined in (19), and $b_{\lambda,\mathcal{O}}$ is defined as:

$$b_{\lambda,\mathcal{O}} = \sum_{(u,v) \in \lambda_p \cap \lambda_r} \log(c\mathbf{1}(u \prec v) + (1 - c)\mathbf{1}(v \prec u))$$

where $u \prec v$ indicates that u precedes v in the ordering $\mathcal{O}$.

D. Application of Priors in Constraint-Based Methods

This section shows the application of priors in constraint-based methods, which typically remove edges by CI tests from a complete undirected graph[9] to derive the output. Due to the lack of a backtracking mechanism, path existence constraints of λ_p are not suited for these methods. For a directed path $(u_0, u_1, \ldots, u_k)$, its edge (u_i, u_{i+1}) cannot be recovered if it is removed before all paths from u_0 to u_k are cut. Therefore, we consider edge absence constraints from λ_r here.

A constraint-based method consists of a a CI test approach $\mathcal{T}$, a strategy $\mathcal{M}$ of processing CI tests among nodes, and an initial (complete undirected) graph G_c:

$$G^{\text{opt}} \leftarrow \mathcal{M}_{\mathcal{G}_n}\left(G_c, D, \mathcal{T}\right) \tag{27}$$

To integrate the edge absence constraint, we set the edge set of the initial graph as the pairs in $\lambda_r \cup \lambda_p$:

$$E(G_{c,r}) = \{(u, v) \mid (u, v) \in \lambda_r \cup \lambda_p\} \tag{28}$$

Since $\mathcal{M}$ does not add new or re-orient directed edges, the edges not present in the initial graph are also absent in the solution. Thus, we have the integration of constraints of λ_r for the constraint-based structure learning:

$$\mathcal{M}_{\mathcal{G}_n}\left(G_c, D, \mathcal{T}\right) \xrightarrow{\lambda_H} \mathcal{M}_{\lambda_r \cup \lambda_p}\left(G_{c,r}, D, \mathcal{T}\right) \tag{29}$$

Reducing the initial graph from G_c to $G_{c,r}$ has a greater impact than merely removing some edges from the output. It can also aid in recovering missing true edges when starting from G_c. Consider an edge (u, v) that is present in the output of $\mathcal{M}$ but removed from G_c. Then u is no longer a neighbor of v and will not be part of the condition set for CI tests. All edges (x, v) where $x \perp\!\!\!\perp v \mid u$ are removed when starting from G_c, while some of them, (x', v), are preserved in this case where $\forall X \subseteq V \setminus \{x', u, v\}$ such that $x' \not\perp\!\!\!\perp v \mid X$. Therefore, edge absence constraints from λ_r can help discovery of new causality

[9]While not all constraint-based methods start from a complete undirected graph, this approach, as adopted in the PC algorithm, is dominant.

Algorithm 5: Gradient-Based Structure Learning.

Require: Data $D \in \mathbb{R}^{m\times n}$, Harmonized Prior λ_r, Objective Function $F(W;D)$, Acyclicity Constraint $h(W)$, Optimization Method $\mathcal{M}$
Ensure: Learned Structure G^{opt}
1: Define Optimization Problem:
$\min_{W_r\in\mathbb{R}^{n\times n}} F(W;D)$
subject to $h(W)=0$
$W_{r,i,j}=0$ for $(x_i,x_j)\in \bar{\lambda_r}\cap\bar{\lambda_p}$
2: $W^{\text{opt}} \leftarrow$ Use $\mathcal{M}$ to solve the equality-constrained problem
3: Construct G^{opt} from W^{opt}
4: **return** G^{opt}

in constraint-based methods. The algorithm of constraint-based structure learning with harmonized prior is presented in Algorithm 4.

E. Application of Priors in Gradient-Based Methods

This section shows the application of priors in gradient-based methods, which solve the following equality-constrained problem to derive outputs:

$$\min_{W\in\mathbb{R}^{n\times n}} F(W;D) \quad \text{subject to } h(W)=0 \tag{30}$$

Here, W is the weighted adjacency matrix whose graph is defined by $(x_i,x_j)\in E(G) \Longleftrightarrow W_{i,j}\neq 0$. The differentiable equality $h(W)=0$ represents the acyclicity of the graph.

Applying path existence constraints in λ_p requires a continuous characterization to capture the existence of paths. Unfortunately, accurately capturing the information of path existence is challenging and remains unresolved. Hence, we only consider the edge absence constraints from λ_r.

To ensure the absence of edges out of $\lambda_r\cup\lambda_p$, their corresponding parameters are frozen to zero:

$$\begin{aligned}&\min_{W_r\in\mathbb{R}^{n\times n}} F(W;D) \quad \text{s.t. } h(W)=0\\ &W_{r,i,j}=0 \text{ for } (x_i,x_j)\in\bar{\lambda_r}\cap\bar{\lambda_p}\end{aligned} \tag{31}$$

Given an optimization method $\mathcal{M}$ that solves the ECP, structure learning with edge absence can be formalized as:

$$\mathcal{M}_{\mathcal{G}_n}(D,W) \xrightarrow{\lambda_H} \mathcal{M}_{\lambda_r\cup\lambda_p}(D,W_r) \tag{32}$$

We show that removing these edges help recover missing edges in this method. Assume that the parent node set of x_i in the optimal solution is Pa_i^{opt}, and we banish the edge (u,x_i) where $u\in\text{Pa}_i^{\text{opt}}$. For the new optimal parent node set[10] $\text{Pa}_i^{\text{opt}'}$ such that $u\notin\text{Pa}_i^{\text{opt}'}$, new edges can be recovered: $\{(u',x_i)\mid u'\in \text{Pa}_i^{\text{opt}'}\setminus\text{Pa}_i^{\text{opt}}\}$. Thus, new causality can be discovered with the help of edge absence constraints of λ_r. The algorithm of gradient-based structure learning with harmonized prior is presented in Algorithm 5.

[10]Assume that this parent node set does not violate the acyclicity constraint and is not the subset of the original optimal one.

TABLE II
OVERVIEW OF STRUCTURE LEARNING METHODS, TYPES OF LLM-DERIVED PRIORS UTILIZED, AND THE INTEGRATION APPROACHES APPLIED

Method	Prior used	Prior constraints	Manner
Score-based	λ_p; λ_r	Ancestral; Edge absence	Hard and soft
Order-based	λ_p; λ_r	Order; Edge Absence	Hard
Constraint-based	λ_r	Edge Absence	Hard
Gradient-based	λ_r	Edge Absence	Hard

Remark 2: The gradient-based algorithm can be seen as a continuous counterpart of the score-based method. In our approach, we enforce the absence of edges in a strict manner by setting the corresponding parameters in the weighted adjacency matrix to zero. However, it is also possible to apply a softer approach in practice. Specifically, one could solve the following optimization problem:

$$\min_W F(W;D)+a\sum_{(x_i,x_j)\notin\lambda_r}|W_{i,j}| \quad \text{s.t. } h(W)=0$$

where $a>0$ represents the weight assigned to the prior term. This prior term allows a balance between the data fitting objective and prior adherence.

F. Discussions on the Impact of LLM-Based Priors

An overview of the structure learning algorithms used in this paper, along with the integration of LLM-derived priors is presented in Table II. This section discusses the influence of these priors on the performance of structure learning.

Causal discovery aims to uncover *new* causal mechanisms from the data, and this discusses in more detail how the LLM-based priors on *known* causality inform this purpose. In structure learning, better discovery of *new* causal mechanisms indicates recovering more accurate structures beyond the *known* structures. The known structures (LLM-based priors) in this paper include existence of paths, absence of edges and order of variables. None of them specifies existence of a specific edge, which can however aid in discovery of edges originally missed in data-based algorithms.

For the path existence constraint in λ_p, it can intuitively help recover missing edges by restoring missing paths. Besides, edge absence constraints from λ_r also contribute to recovering missing edges. In constraint-based methods, this is achieved by reducing the condition set of CI tests to preserve missing edges, as discussed in Section IV-D. In score-based and gradient-based methods, λ_r constraints assist by reducing the hypothesis space, as detailed in Section IV-E.

Using λ_p as order constraints can help recover missing edges by reducing the order space. Suppose an edge (u,v) is in the optimal solution, and we have an order constraint $v\prec u$. This constraint eliminates all orderings that include the optimal solution, forcing a new optimal solution with a different variable ordering. Consequently, this changes the graph space, as candidate parent nodes of variables vary with different orderings. Thus, missing edges can be recovered.

The aspect of LLM's impact on discovery of new causality can be represented by the differences of true positive rate (TPR),

TABLE III
DETAILS OF THE REAL-WORLD STRUCTURES USED

Dataset	Nodes	Arcs	Paras	Research domain
Cancer	5	4	10	Smoking and cancer
Asia	8	8	18	Reasons causing dyspnoea
Child	20	25	230	Childhood diseases
Insurance	27	52	1008	Car insurance cost
Mildew	35	46	540k	Winter wheat mildew control
Alarm	37	46	509	ALARM monitoring system

Columns nodes, arcs and paras represent the number of nodes and edges of the structure, and parameters of the graphical model, respectively.

which is taken as a critical evaluation metric in the following experiments.

V. EXPERIMENTS

This section evaluates the proposed LLM-driven structure learning framework. We begin by reporting the alignment of LLM-derived priors with real causal structures to demonstrate the rationale behind the prompting strategies and structural constraints used. Next, we compare the output quality of our LLM-driven approach with data-based methods to highlight the performance improvements. We then conduct ablation studies on the harmonized prior to showcase the effectiveness of integrating both priors, λ_p and λ_r. Following this, we delve into the soft constraint approach for deeper insights of its error-tolerant ability. Finally, we assess the impact on time efficiency by integrating LLM-derived priors to evaluate its practical feasibility.

A. Datasets and Setup

We utilize six real-world structures from the BN repository.[11] The details of these structures and the context of their variables are presented in Table III. For BN structure learning, the observational data is sourced from the study by Li et al. [55], encompassing two sizes of samples for each structure and six simulations for each size. For linear SEM structure learning, we generate synthetic data based on the linear SEM with a Gaussian noise term. For nonlinear SEM structure learning, we generate data using the Index model [68]. For a structure with n nodes, we generate four different sample sizes, kn, where $k \in \{1, 2, 4, 10\}$, with six simulations for each size.

We implement the LLM-driven structure learning framework across mainstream structure learning algorithms, with corresponding details provided in Table IV. We examine using Claude-2 [69], GPT-3.5-turbo and GPT-4 [70] to derive priors and report their qualities. The priors from GPT-4 are applied in the data-based structure learning process. All tests run on an AMD Ryzen9 7950x CPU at 4.5 GHz, with a 16 GB memory cap and a 24-hour time limit.

B. Metrics

We use three structural metrics to evaluate the quality of output structures: Structural Hamming Distance (SHD), True Positive Rate (TPR), and F1-score.

[11] https://www.bnlearn.com/bnrepository/

SHD measures the number of differing directed edges between the output structure and the ground truth, making it a widely used structural metric. However, SHD alone may not comprehensively reflect the quality of structures.[12]

Therefore, we also use TPR and F1-score. TPR represents the recall rate of edges in the output structure, while F1-score provides a balanced evaluation of the recall and precision rates of edges in the output structure.

C. Quality of LLM-Derived Prior

This experiment evaluates the quality of priors derived from various LLMs, including Claude-2 [69], GPT-3.5-turbo and GPT-4 [70]. For the precision-oriented prior λ_p, we report accuracies based on consistency with the existence of an edge, a directed path, and order.[13] For the recall-oriented prior λ_r, we report the accuracy of using its complementary set to indicate edge absence. The results are shown in Table V.

Our observations indicate that the accuracy of λ_p improves for looser structural constraints due to the derivation chain $x_i \to x_j$ (Edge) $\Rightarrow x_i \rightsquigarrow x_j$ (Path) $\Rightarrow x_i \prec x_j$ (Order). For GPT-4, the most advanced LLM in our study, we see a marked decrease in accuracy from path to edge. This finding supports our use of λ_p as path existence constraints in structure learning. In certain datasets, such as *Child* and *Mildew*, path accuracy is less satisfactory. This occurs because many seemingly correct cause-and-effect pairs are actually connected by a confounder, where the path does not exist, but the order is still correct. This also explains why order constraints maintain higher accuracy in these cases.

In contrast, the accuracy of edge absence using λ_r from GPT-4 remains consistently high across all datasets, demonstrating the effectiveness of our specialized prompting strategies in guiding LLM behavior.

D. Comparison to Data-Based Structure Learning

In this experiment, we compare the output quality of our LLM-driven structure learning methods with their data-based counterparts. For each data-based method m, we denote its LLM-driven version using the harmonized prior as H-m. Note that various methods can use different versions of the harmonized prior, as detailed in Table IV. The results for methods on BN data are reported in Tables VI, VII, and VIII, with two datasets in each table. Results for methods on SEM data are reported in Tables IX and X.

We observe that on *Asia*, *Cancer*, and *Insurance* datasets, which have high-quality LLM-derived priors, the LLM-driven methods consistently outperform their data-based counterparts, as evidenced by enhanced TPR or F1 scores in each case. The improvement in TPR, in particular, highlights the benefit of integrating known causal knowledge to aid in recovering better unknown causal mechanisms, as discussed in Section IV-F. Notably, on the *Cancer* dataset with 500 samples, H-Tabu and

[12] For example, an empty graph has an SHD equal to the edge count, which can sometimes appear better than some state-of-the-art algorithms in certain cases.

[13] The pair (u, v) does not violate the order of u and v if v is not an ancestor of u in the true causal graph.

TABLE IV
DETAILS OF THE STRUCTURE LEARNING ALGORITHMS USED

DAG Model	Type	Algorithm	Score	Prior	Integration
BN	Score-based	Tabu [64]	BIC, BDeu	λ_H (path existence for λ_p)	Hard
		MCMC [65]	MML	λ_H (path existence for λ_p)	Soft
	Order-based	A* [56]	BIC, BDeu	λ_H (order constraint for λ_p)	Hard
		DP [66]	BIC, BDeu	λ_H (order constraint for λ_p)	Hard
	Constraint-based	PC [21]	–	λ_r ($E(G) \subseteq \lambda_r \cup \lambda_p$)	Hard
Linear SEM	Gradient-based	NOTEARS [46]	Least square	λ_r ($E(G) \subseteq \lambda_r \cup \lambda_p$)	Hard
Nonlinear SEM	Gradient-based	NOTEARS-MLP [67]	Least square	λ_r ($E(G) \subseteq \lambda_r \cup \lambda_p$)	Hard

Note that order-based search is a subset of score-based methods, distinguished here due to different applications of priors. Additionally, Tabu and MCMC are approximate searches, while A* and DP are exact searches. The symbol λ_p (path) indicates the use of λ_p as path existence, and λ_p (order) indicates its use as order constraints.

TABLE V
CONSISTENCY BETWEEN THE DERIVED PRIOR CONSTRAINTS BY VARIOUS LLMS WITH THE TRUE CAUSAL STRUCTURES

LLM	Dataset	λ_p			λ_r
		@Edge	@Path	@Order	@EdgeAb
GPT-4	Cancer	0.80	1.00	1.00	1.00
	Asia	0.78	1.00	1.00	1.00
	Alarm	0.62	0.95	0.95	0.99
	Child	0.50	0.64	0.93	1.00
	Mildew	0.33	0.67	0.89	0.99
	Insurance	0.70	1.00	1.00	1.00
GPT-3.5-turbo	Cancer	1.00	1.00	1.00	0.93
	Asia	0.57	0.86	0.86	0.93
	Alarm	0.39	0.65	1.00	0.98
	Child	0.55	0.64	0.91	0.97
	Mildew	0.60	0.60	0.88	0.99
	Insurance	0.17	0.67	1.00	0.95
Claude-2	Cancer	1.00	1.00	1.00	1.00
	Asia	1.00	1.00	1.00	0.93
	Alarm	0.55	0.69	0.90	0.98
	Child	0.20	0.33	0.73	0.98
	Mildew	0.88	0.91	0.91	0.98
	Insurance	0.63	0.88	1.00	0.95

Columns '@Edge', '@Path' and '@Order' is the accuracy of priors w.r.t. existence of edges, paths and orders, respectively. Column '@EdgeAb' is the accuracy of the complement set of priors w.r.t. absence of edges.

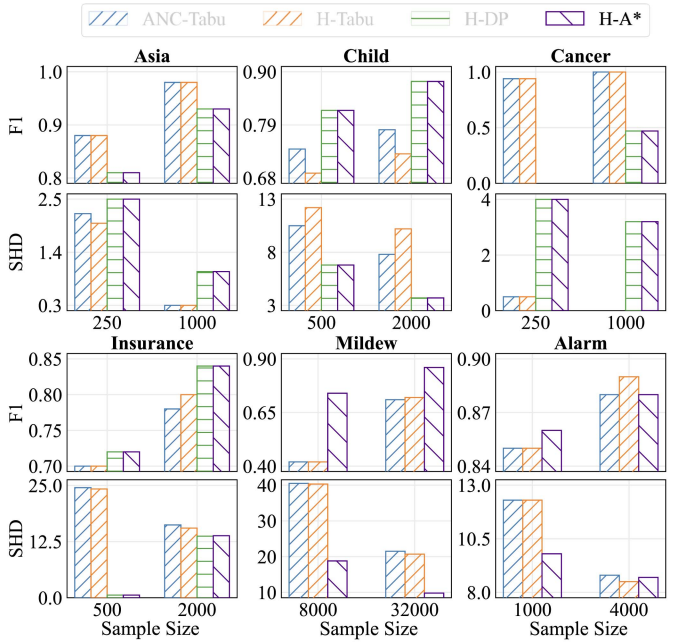


Fig. 4. Comparison of our method and the previous LLM-driven structure learning approach in terms of F1 score and SHD.

H-MCMC show improvement, while H-A*, H-DP, and H-PC exhibit comparable performance to their data-based counterparts. This difference is because the latter three methods do not ensure the existence of paths in λ_p.

On the *Child* and *Mildew* datasets, which have lower-quality LLM-derived priors, we observe that the Tabu method shows degraded performance for its LLM-driven version, while other methods still show improvement. The reason for Tabu's degraded performance is that using λ_p as path existence is not accurate on these datasets. In contrast, A* and DP maintain their enhancements as they use λ_p as order constraints, which are accurate for these datasets. This demonstrates that using λ_p as order constraints, looser than path existence, can lead to stable performance even with inaccurate LLM knowledge.

Apart from H-A* and H-DP, H-PC and H-NOTEARS(-MLP) also show improvements over their data-based counterparts in almost all cases, as these methods only employ edge absence of λ_r, which is accurate across all datasets. An interesting case is H-MCMC, which uses λ_p as path existence like H-Tabu but still improves over MCMC in nearly all cases. This stability stems from its soft approach to integrating structural constraints, which is further investigated later.

E. Comparison to LLM-Driven Structure Learning

In this experiment, we compare our LLM-driven structure learning approach to the previous method proposed in [9], referred to as ANC-*method*, which uses LLM-derived priors as ancestral constraints. Using the BDeu score as the scoring function, we evaluate and compare H-Tabu, H-DP, and H-A* against ANC-Tabu, with results presented in Fig. 4.

The results indicate that our approach outperforms the previous LLM-driven method in most scenarios. This improvement stems from our method's ability to derive more precise priors through a fine-grained prompting strategy, alongside the integration of additional edge absence constraints, which together

TABLE VI
RESULTS OF METHODS BASED ON THE BN ON THE ASIA AND CHILD DATASETS

Dataset		Asia						Child					
Size		250			1000			500			2000		
Score	Method	SHD	TPR	F1	SHD	TPR	F1	SHD	TPR	F1	SHD	TPR	F1
–	PC	1.3 ± 1.2	0.90 ± 0.09	0.62 ± 0.05	2.2 ± 0.8	0.94 ± 0.07	0.57 ± 0.03	32.5 ± 2.9	0.51 ± 0.05	0.28 ± 0.03	35.8 ± 2.0	0.69 ± 0.02	0.30 ± 0.01
	H-PC	1.3 ± 0.5	1.00 ± 0.00	0.79 ± 0.02	2.5 ± 1.0	1.00 ± 0.00	0.72 ± 0.05	27.3 ± 4.1	0.56 ± 0.07	0.40 ± 0.05	26.2 ± 2.2	0.71 ± 0.02	0.43 ± 0.02
BDeu	A*	3.8 ± 1.2	0.56 ± 0.10	0.66 ± 0.11	2.0 ± 0.9	0.77 ± 0.09	0.82 ± 0.10	7.3 ± 1.4	0.71 ± 0.05	0.80 ± 0.04	4.2 ± 2.1	0.83 ± 0.09	0.87 ± 0.08
	H-A*	2.5 ± 0.5	0.69 ± 0.07	0.81 ± 0.05	1.0 ± 0.0	0.88 ± 0.00	0.93 ± 0.00	6.8 ± 0.4	0.73 ± 0.02	0.82 ± 0.01	3.7 ± 0.8	0.85 ± 0.03	0.88 ± 0.02
	DP	3.7 ± 1.4	0.58 ± 0.13	0.69 ± 0.14	2.2 ± 0.8	0.75 ± 0.08	0.80 ± 0.08	8.2 ± 2.0	0.68 ± 0.07	0.76 ± 0.07	5.0 ± 1.4	0.80 ± 0.06	0.83 ± 0.06
	H-DP	2.5 ± 0.5	0.69 ± 0.07	0.81 ± 0.05	1.0 ± 0.0	0.88 ± 0.00	0.93 ± 0.00	6.8 ± 0.4	0.73 ± 0.02	0.82 ± 0.01	3.7 ± 0.8	0.85 ± 0.03	0.88 ± 0.02
	Tabu	4.2 ± 2.6	0.69 ± 0.21	0.68 ± 0.22	2.5 ± 0.5	0.73 ± 0.05	0.75 ± 0.07	9.5 ± 2.1	0.66 ± 0.07	0.72 ± 0.07	5.3 ± 1.0	0.79 ± 0.05	0.82 ± 0.03
	H-Tabu	2.0 ± 1.3	0.94 ± 0.07	0.88 ± 0.07	0.3 ± 0.5	1.00 ± 0.00	0.98 ± 0.03	12.2 ± 1.7	0.67 ± 0.05	0.69 ± 0.04	10.2 ± 1.2	0.76 ± 0.05	0.73 ± 0.04
BIC	A*	3.7 ± 2.3	0.67 ± 0.17	0.71 ± 0.19	1.3 ± 0.8	0.85 ± 0.05	0.90 ± 0.05	7.8 ± 1.5	0.71 ± 0.06	0.78 ± 0.06	3.0 ± 0.6	0.88 ± 0.03	0.89 ± 0.02
	H-A*	1.5 ± 0.8	0.83 ± 0.10	0.90 ± 0.07	0.8 ± 0.4	0.90 ± 0.05	0.94 ± 0.03	6.3 ± 0.5	0.75 ± 0.02	0.83 ± 0.01	2.7 ± 0.5	0.89 ± 0.02	0.91 ± 0.01
	DP	3.5 ± 2.1	0.69 ± 0.13	0.73 ± 0.15	1.2 ± 1.0	0.88 ± 0.08	0.92 ± 0.07	8.2 ± 1.3	0.70 ± 0.05	0.77 ± 0.06	3.0 ± 0.9	0.88 ± 0.04	0.89 ± 0.03
	H-DP	1.5 ± 0.8	0.83 ± 0.10	0.90 ± 0.07	0.8 ± 0.4	0.90 ± 0.05	0.94 ± 0.03	6.2 ± 0.4	0.75 ± 0.02	0.83 ± 0.01	2.7 ± 0.5	0.89 ± 0.02	0.91 ± 0.01
	Tabu	2.2 ± 1.5	0.83 ± 0.10	0.86 ± 0.10	0.8 ± 0.8	0.92 ± 0.06	0.95 ± 0.05	8.5 ± 1.5	0.69 ± 0.05	0.75 ± 0.05	4.5 ± 1.0	0.82 ± 0.04	0.83 ± 0.04
	H-Tabu	1.0 ± 1.5	0.98 ± 0.05	0.94 ± 0.09	0.2 ± 0.4	1.00 ± 0.00	0.99 ± 0.02	13.0 ± 2.4	0.64 ± 0.06	0.67 ± 0.05	9.0 ± 1.1	0.82 ± 0.07	0.77 ± 0.04
MML	MCMC	3.7 ± 1.2	0.60 ± 0.10	0.66 ± 0.11	2.2 ± 0.4	0.75 ± 0.00	0.78 ± 0.02	**5.6 ± 1.6**	0.78 ± 0.06	**0.84 ± 0.05**	**1.7 ± 1.0**	0.93 ± 0.04	0.93 ± 0.04
	H-MCMC	**0.0 ± 0.0**	**1.00 ± 0.00**	**1.00 ± 0.00**	**0.0 ± 0.0**	**1.00 ± 0.00**	**1.00 ± 0.00**	6.8 ± 1.0	**0.81 ± 0.05**	0.82 ± 0.03	1.8 ± 0.5	**0.96 ± 0.00**	**0.95 ± 0.01**

Gray text indicates the worse performance between a method *m* and its LLM-driven counterpart H-*m*. The best performance for each dataset is highlighted in bold.

TABLE VII
RESULTS OF METHODS BASED ON THE BN ON CANCER AND INSURANCE DATASETS

Dataset		Cancer						Insurance					
Size		250			1000			500			2000		
Score	Method	SHD	TPR	F1	SHD	TPR	F1	SHD	TPR	F1	SHD	TPR	F1
–	PC	1.0 ± 0.0	0.75 ± 0.00	0.60 ± 0.00	0.8 ± 0.4	0.79 ± 0.10	0.61 ± 0.03	64.7 ± 3.3	0.29 ± 0.03	0.22 ± 0.02	63.3 ± 1.8	0.43 ± 0.02	0.27 ± 0.01
	H-PC	1.0 ± 0.0	0.75 ± 0.00	0.75 ± 0.00	0.8 ± 0.4	0.79 ± 0.10	0.75 ± 0.01	73.5 ± 1.8	0.35 ± 0.01	0.26 ± 0.01	69.8 ± 1.9	0.50 ± 0.02	0.31 ± 0.01
BDeu	A*	4.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00	3.3 ± 0.5	0.17 ± 0.13	0.38 ± 0.03	25.2 ± 2.2	0.55 ± 0.03	0.67 ± 0.04	16.5 ± 2.3	0.69 ± 0.03	0.78 ± 0.03
	H-A*	4.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00	3.2 ± 0.8	0.21 ± 0.19	0.47 ± 0.13	21.7 ± 1.5	0.59 ± 0.02	0.72 ± 0.02	13.8 ± 1.7	**0.74 ± 0.03**	**0.84 ± 0.03**
	DP	4.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00	3.3 ± 0.5	0.17 ± 0.13	0.38 ± 0.03	–	–	–	–	–	–
	H-DP	4.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00	3.2 ± 0.8	0.21 ± 0.19	0.47 ± 0.13	21.7 ± 1.5	0.59 ± 0.02	0.72 ± 0.02	**13.7 ± 1.0**	**0.74 ± 0.02**	**0.84 ± 0.02**
	Tabu	3.0 ± 0.9	0.29 ± 0.25	0.58 ± 0.13	1.8 ± 1.2	0.54 ± 0.29	0.60 ± 0.29	25.7 ± 2.3	0.58 ± 0.04	0.68 ± 0.04	15.0 ± 2.1	**0.74 ± 0.03**	0.81 ± 0.04
	H-Tabu	**0.5 ± 0.8**	**0.96 ± 0.10**	**0.94 ± 0.10**	**0.0 ± 0.0**	**1.00 ± 0.00**	**1.00 ± 0.00**	24.2 ± 2.3	0.60 ± 0.02	0.70 ± 0.03	15.5 ± 1.9	0.73 ± 0.03	0.80 ± 0.03
BIC	A*	3.8 ± 0.8	0.08 ± 0.13	0.40 ± 0.00	2.5 ± 0.5	0.38 ± 0.14	0.45 ± 0.11	43.7 ± 12.3	0.42 ± 0.05	0.47 ± 0.08	31.8 ± 7.0	0.51 ± 0.03	0.56 ± 0.04
	H-A*	3.8 ± 0.8	0.08 ± 0.13	0.40 ± 0.00	1.7 ± 0.8	0.58 ± 0.20	0.72 ± 0.18	28.2 ± 3.4	0.53 ± 0.05	0.63 ± 0.04	16.5 ± 1.8	0.71 ± 0.02	0.81 ± 0.02
	DP	3.8 ± 0.8	0.08 ± 0.13	0.40 ± 0.00	2.0 ± 0.9	0.50 ± 0.22	0.60 ± 0.20	–	–	–	–	–	–
	H-DP	3.8 ± 0.8	0.08 ± 0.13	0.40 ± 0.00	1.7 ± 0.8	0.58 ± 0.20	0.72 ± 0.18	25.7 ± 2.5	0.54 ± 0.05	0.66 ± 0.04	15.5 ± 0.8	0.71 ± 0.03	0.82 ± 0.01
	Tabu	2.2 ± 0.8	0.50 ± 0.16	0.64 ± 0.15	0.8 ± 0.8	0.79 ± 0.19	0.87 ± 0.12	33.2 ± 2.9	0.41 ± 0.04	0.52 ± 0.06	29.5 ± 2.1	0.49 ± 0.04	0.56 ± 0.04
	H-Tabu	3.0 ± 1.1	0.63 ± 0.14	0.63 ± 0.14	1.0 ± 1.1	0.88 ± 0.14	0.88 ± 0.14	31.0 ± 2.3	0.49 ± 0.03	0.60 ± 0.04	21.2 ± 2.4	0.64 ± 0.03	0.72 ± 0.03
MML	MCMC	3.7 ± 0.7	0.08 ± 0.16	0.44 ± 0.19	2.3 ± 0.5	0.42 ± 0.12	0.48 ± 0.11	35.3 ± 1.6	0.39 ± 0.03	0.46 ± 0.04	32.6 ± 4.3	0.48 ± 0.06	0.53 ± 0.06
	H-MCMC	2.0 ± 1.2	0.75 ± 0.15	0.75 ± 0.15	0.3 ± 0.8	0.96 ± 0.10	0.96 ± 0.10	**20.8 ± 2.3**	**0.63 ± 0.04**	**0.73 ± 0.04**	20.5 ± 4.6	0.66 ± 0.08	0.73 ± 0.07

The symbol '–' indicates that the method failed to complete within the 16GB memory or 24-hour time limit.

contribute to greater robustness and overall effectiveness in structure learning.

Notably, H-A* and H-DP exhibit the best performance on larger datasets (those out of *Asia* and *Cancer*) due to the use of order constraints, which has increased effectiveness in reducing the search space as the number of variables grows and is robust against prior errors. The use of order constraints to manage larger variable sets is an unexplored aspect in the previous study [9], highlighting the strength of our approach in addressing more complex scenarios.

F. Ablation Study of Harmonized Prior

This experiment investigates the individual impact of the two aspects of LLM-derived priors, λ_p and λ_r, on structure learning. We report on three methods, each using a distinct strategy on λ: A* uses order constraints, Tabu employs hard constraints on path existence, and MCMC utilizes soft constraints on path existence. Results for TPR, F1, and SHD on the datasets that all these methods successfully complete are reported in Fig. 5.

We observe that integrating both λ_p and λ_r enhances performance more than using each prior individually. This is evidenced by cases showing increased performance when using both priors, such as all metrics for MCMC on *Cancer*, F1 and SHD for Tabu on *Asia*, TPR for MCMC on *Child*, and all metrics for A* and MCMC on *Insurance*. Moreover, the integration of both priors introduces more stable improvements in data-based methods compared to using a single prior. For example, using only λ_p leads to more nuanced performance, while λ_r results in

TABLE VIII
RESULTS OF METHODS BASED ON THE BN ON MILDEW AND ALARM DATASETS

Dataset		Mildew						Alarm					
Size		8000			32000			1000			4000		
Score	Method	SHD	TPR	F1	SHD	TPR	F1	SHD	TPR	F1	SHD	TPR	F1
–	PC	71.2 ± 2.9	0.75 ± 0.04	0.29 ± 0.01	63.0 ± 3.8	0.81 ± 0.03	0.33 ± 0.01	66.8 ± 3.7	0.47 ± 0.02	0.25 ± 0.01	84.0 ± 3.4	0.56 ± 0.01	0.23 ± 0.01
	H-PC	72.0 ± 2.0	**0.90 ± 0.02**	0.40 ± 0.01	65.5 ± 2.2	**0.94 ± 0.01**	0.44 ± 0.01	76.5 ± 4.2	0.54 ± 0.03	0.28 ± 0.02	91.0 ± 1.8	0.63 ± 0.01	0.26 ± 0.01
BDeu	A*	–	–	–	–	–	–	–	–	–	–	–	–
	H-A*	**18.8 ± 0.4**	0.63 ± 0.01	**0.74 ± 0.01**	**9.8 ± 0.4**	0.79 ± 0.01	**0.86 ± 0.01**	9.8 ± 1.3	0.85 ± 0.03	0.86 ± 0.03	8.7 ± 0.5	0.88 ± 0.01	0.88 ± 0.01
	Tabu	22.8 ± 0.8	0.55 ± 0.02	0.63 ± 0.02	21.0 ± 2.2	0.67 ± 0.02	0.67 ± 0.03	9.5 ± 2.7	**0.91 ± 0.04**	0.88 ± 0.05	6.5 ± 1.8	0.92 ± 0.03	0.91 ± 0.02
	H-Tabu	40.3 ± 1.8	0.40 ± 0.02	0.42 ± 0.02	20.7 ± 3.1	0.72 ± 0.06	0.72 ± 0.06	12.3 ± 2.3	0.89 ± 0.03	0.85 ± 0.03	8.5 ± 1.5	0.91 ± 0.01	0.89 ± 0.02
BIC	A*	–	–	–	–	–	–	–	–	–	–	–	–
	H-A*	28.3 ± 1.2	0.49 ± 0.02	0.60 ± 0.02	20.0 ± 0.0	0.61 ± 0.00	0.72 ± 0.00	12.5 ± 1.8	0.78 ± 0.03	0.81 ± 0.03	10.5 ± 0.5	0.84 ± 0.02	0.85 ± 0.01
	Tabu	34.2 ± 0.4	0.37 ± 0.01	0.43 ± 0.01	33.5 ± 4.0	0.48 ± 0.06	0.50 ± 0.07	14.8 ± 3.8	0.72 ± 0.06	0.75 ± 0.07	6.8 ± 1.8	0.86 ± 0.03	0.88 ± 0.04
	H-Tabu	45.8 ± 1.5	0.34 ± 0.02	0.36 ± 0.02	41.8 ± 2.7	0.42 ± 0.03	0.43 ± 0.03	14.5 ± 2.2	0.78 ± 0.04	0.79 ± 0.04	9.2 ± 0.8	0.87 ± 0.02	0.87 ± 0.01
MML	MCMC	49.2 ± 3.2	0.26 ± 0.04	0.26 ± 0.04	59.9 ± 4.2	0.27 ± 0.08	0.23 ± 0.07	11.4 ± 2.1	0.82 ± 0.03	0.82 ± 0.03	9.3 ± 3.3	0.85 ± 0.04	0.85 ± 0.05
	H-MCMC	48.9 ± 2.5	0.26 ± 0.04	0.27 ± 0.04	59.5 ± 6.0	0.28 ± 0.11	0.24 ± 0.09	**6.0 ± 3.2**	**0.91 ± 0.05**	**0.92 ± 0.05**	**3.3 ± 2.0**	**0.95 ± 0.02**	**0.96 ± 0.03**

Results for methods DP and H-DP are omitted as both failed to complete on these datasets within the 16GB memory limit.

TABLE IX
RESULTS OF THE METHOD BASED ON THE LINEAR SEM (NOTEARS) ON ALL DATASETS

Metric	Method	Cancer				Asia				Child			
		1	2	4	10	1	2	4	10	1	2	4	10
SHD	NOTEARS	6.3 ± 1.5	4.2 ± 1.9	3.0 ± 2.4	0.8 ± 1.2	14.8 ± 3.1	7.7 ± 2.2	3.3 ± 1.6	0.3 ± 0.5	55.0 ± 8.2	17.2 ± 1.7	8.7 ± 4.4	4.0 ± 4.5
	H-NOTEARS	**5.2 ± 2.3**	**3.7 ± 2.0**	**2.7 ± 2.8**	**0.8 ± 1.2**	**5.3 ± 2.6**	**2.5 ± 1.4**	**1.0 ± 1.5**	**0.0 ± 0.0**	**20.8 ± 6.1**	**6.5 ± 2.4**	**2.8 ± 1.0**	**0.7 ± 0.5**
F1	NOTEARS	.33 ± .18	.56 ± .19	.69 ± .26	.92 ± .11	.35 ± .14	.62 ± .14	.80 ± .10	.98 ± .03	.38 ± .07	.73 ± .01	.83 ± .08	.91 ± .10
	H-NOTEARS	**.49 ± .24**	**.61 ± .22**	**.73 ± .29**	**.92 ± .11**	**.72 ± .15**	**.87 ± .07**	**.95 ± .08**	**1.0 ± .00**	**.69 ± .08**	**.89 ± .04**	**.93 ± .02**	**.97 ± .02**

Metric	Method	Insurance				Mildew				Alarm			
		1	2	4	10	1	2	4	10	1	2	4	10
SHD	NOTEARS	63.0 ± 8.9	34.2 ± 9.8	22.3 ± 10.9	27.8 ± 6.9	91.8 ± 9.5	49.0 ± 7.8	42.7 ± 8.1	39.2 ± 4.2	88.0 ± 7.2	32.5 ± 13.3	13.3 ± 11.6	19.7 ± 12.6
	H-NOTEARS	**22.5 ± 2.7**	**6.8 ± 3.1**	**2.2 ± 2.5**	**0.3 ± 0.8**	**20.2 ± 6.0**	**3.8± 2.2**	**0.0 ± 0.0**	**0.0 ± 0.0**	**29.3 ± 12.7**	**7.2 ± 4.5**	**1.5 ± 1.4**	**1.5 ± 1.6**
F1	NOTEARS	.51 ± .05	.66 ± .08	.77 ± .11	.72 ± .06	.37 ± .04	.55 ± .05	.57 ± .06	.60 ± .03	.42 ± .04	.67 ± .12	.86 ± .12	.79 ± .13
	H-NOTEARS	**.81 ± .02**	**.93 ± .03**	**.98 ± .03**	**1.0 ± .01**	**.82 ± .05**	**.96 ± .02**	**1.0 ± .00**	**1.0 ± .00**	**.74 ± .09**	**.92 ± .05**	**.98 ± .02**	**.98 ± .02**

The better performance between NOTEARS and H-NOTEARS is highlighted in bold.

TABLE X
RESULTS OF THE METHOD BASED ON THE NONLINEAR SEM (NOTEARS-MLP) ON ALL DATASETS

Metric	Method	Cancer				Asia				Child			
		1	2	4	10	1	2	4	10	1	2	4	10
SHD	NOTEARS-MLP	8.2 ± 0.9	6.2 ± 1.1	6.7 ± 1.5	4.0 ± 1.9	18.3 ± 2.7	18.0 ± 1.8	17.2 ± 0.4	10.7 ± 2.1	118 ± 9.5	126 ± 10.7	112 ± 10.5	39.7 ± 8.9
	H-NOTEARS-MLP	**6.0 ± 1.0**	**6.0 ± 1.2**	**6.2 ± 1.1**	**4.0 ± 1.5**	**9.0 ± 1.7**	**9.0 ± 1.6**	**6.7 ± 0.7**	**5.3 ± 0.9**	**55.7 ± 5.0**	**64.0 ± 5.8**	**59.0 ± 5.4**	**32.2 ± 3.4**
F1	NOTEARS-MLP	.21 ± .07	.41 ± .13	.44 ± .09	.64 ± .17	.29 ± .10	.36 ± .10	.40 ± .04	.51 ± .09	.20 ± .02	.21 ± .03	.25 ± .03	.51 ± .06
	H-NOTEARS-MLP	**.38 ± .08**	**.43 ± .11**	**.44 ± .08**	**.64 ± .09**	**.48 ± .11**	**.54 ± .09**	**.68 ± .04**	**.70 ± .08**	**.25 ± .03**	**.30 ± .04**	**.36 ± .05**	**.53 ± .04**

Metric	Method	Insurance				Mildew				Alarm			
		1	2	4	10	1	2	4	10	1	2	4	10
SHD	NOTEARS-MLP	199 ± 10.7	209 ± 13.6	202 ± 14.1	64.5 ± 17.3	296 ± 7.1	317 ± 14.8	257 ± 27.4	37.2 ± 15.7	314 ± 48.8	340 ± 10.9	247 ± 22.6	40.8 ± 10.9
	H-NOTEARS-MLP	**90.7 ± 6.3**	**103 ± 9.9**	**98.8 ± 5.3**	**43.0 ± 7.5**	**100 ± 9.9**	**109 ± 10.1**	**108 ± 6.8**	**30.3 ± 7.5**	**113 ± 4.5**	**140 ± 6.9**	**115 ± 8.9**	**25.5 ± 2.8**
F1	NOTEARS-MLP	.20 ± .01	.24 ± .03	.27 ± .03	.53 ± .04	.13 ± .01	.16 ± .01	.20 ± .03	.41 ± .01	.14 ± .01	.15 ± .01	.22 ± .02	.65 ± .07
	H-NOTEARS-MLP	**.25 ± .03**	**.32 ± .05**	**.38 ± .03**	**.66 ± .04**	**.32 ± .05**	**.39 ± .04**	**.41 ± .01**	**.72 ± .05**	**.24 ± .02**	**.29 ± .03**	**.38 ± .03**	**.75 ± .02**

The better performance between NOTEARSMLP and H-NOTEARS-MLP is highlighted in bold.

comparable or worse performance than the data-based method, as seen with A* on *Cancer* and MCMC on *Asia*. Conversely, some cases indicate the opposite trend, such as A* on *Child* and *Insurance*. In these cases, integrating both priors still leads to nuanced performance, consistently improving the data-based method.

In cross-method comparisons, we observe that A* and MCMC balance the two priors effectively, whereas Tabu's performance

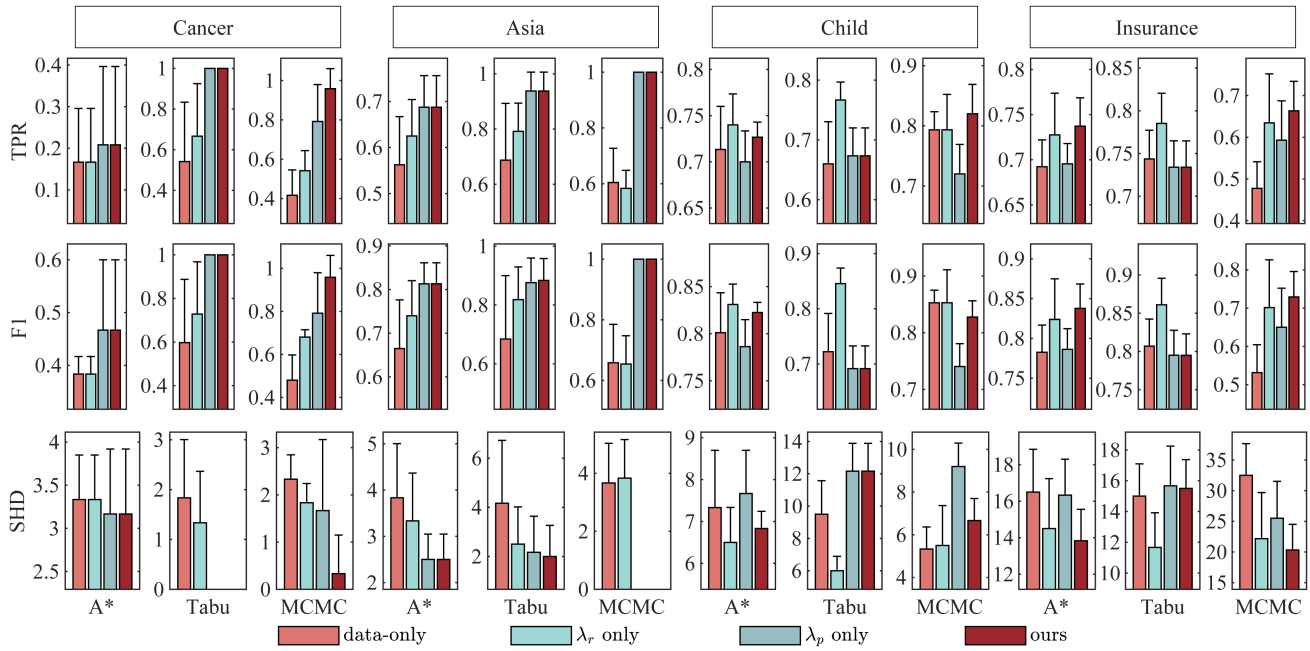


Fig. 5. Ablation study with results of TPR, F1 and SHD on structure learning using only data, only recall-oriented prior λ_r, only precision-oriented prior λ_p, and the harmonized prior (both λ_r and λ_p). The reported results use algorithms A* (with order constraints of λ_p) and Tabu (with hard constraints of path existence of λ_p) with BDeu score, and MCMC (with soft constraints of path existence of λ_p) with MML score. The sample size of the reported results is 1000 for Cancer, 250 for Asia, 500 for Child and 2000 for Insurance.

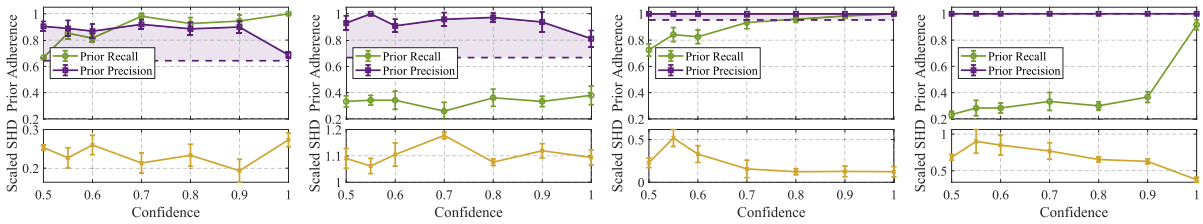


Fig. 6. The precision and recall of accepted priors in λ_p and scaled SHD by MCMC with varying prior probability. The dotted line represents the precision of λ_p from GPT-4. The colored area represents the improvement of the precision of accepted priors in λ_p.

suffers when λ_p leads to worse outcomes than data-based methods, as seen in the *Child* and *Insurance* cases. This is because Tabu uses a hard approach for λ_p, and the extra edges caused by inaccurate path existence constraints cannot be mitigated by incorporating edge absence constraints from λ_r. In contrast, A* and MCMC have error-tolerance mechanisms for λ_p, allowing some inaccuracies to be corrected, thus minimizing harm to the output and balancing with λ_r, which reduces the hypothesis space. The error-tolerance ability of A* stems from a looser but hard use of λ_p, while that of MCMC comes from a soft use of λ_p, controlled by the setting of prior probability. The impact of this prior probability on MCMC's error-tolerance ability is further investigated subsequently.

G. Impact of Prior Probability on the Error-Tolerant Ability

We investigate the error-tolerant ability of the soft application approach, MCMC, with various settings of prior probability on the *Child* (500 samples), *Alarm* (1000 samples), and *Mildew* (8000 samples) datasets, where there are erroneous constraints in λ_p, and on the *Insurance* (500 samples) dataset without prior errors for comparison. We set prior probability from 0.5 (no prior) to 0.99999 (high confidence), assessing precision, recall of accepted constraints, and SHD. To unify the range of SHD values across different datasets, we scale SHD by dividing it by the number of edges in the ground truth. The results are reported in Fig. 6.

The results show that the soft method consistently accepts constraints with higher precision than their inherent accuracy, demonstrating its robustness to prior errors. In the error-free *Insurance* dataset, structure quality improves with increasing prior probability due to the acceptance of more correct constraints. Conversely, in datasets with prior errors, quality can decline at higher prior probability, which enforce adherence to more incorrect constraints.

Therefore, prior probability should be set to a level that balances the benefits of accepting more constraints against the risks of potential errors. For example, setting the prior probability

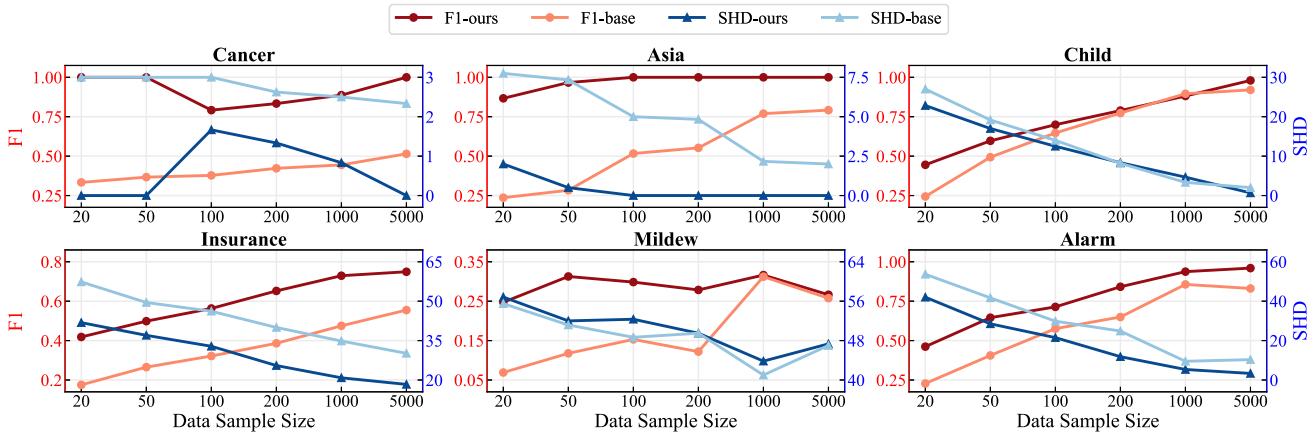


Fig. 7. Comparison of the LLM-driven method via harmonized prior (denoted as "ours") versus the data-based algorithm with no prior (denoted as "base") in terms of F1↑ score and SHD↓ in the case of various data samples. The algorithm used is MCMC.

TABLE XI
COMPARISON IN RUNTIME (IN SECONDS) BETWEEN LLM-DRIVEN AND DATA-BASED STRUCTURE LEARNING

Dataset	Size	Exact search				Approximate search			
		DP		A*		Tabu		MCMC	
		t_{data}	t_{LLM}	t_{data}	t_{LLM}	t_{data}	t_{LLM}	t_{data}	t_{LLM}
Cancer	250	0.8	0.7	0.8	0.9	4.0	4.0	0.7	0.9
	1000	0.8	0.8	0.8	0.9	4.0	4.0	0.7	0.9
Asia	250	1.3	0.7	1.5	0.8	3.6	3.5	1.4	1.4
	1000	1.4	0.8	1.7	0.8	3.7	3.5	1.3	1.4
Child	500	13 k	17.7	0.1 k	11.2	3.5	5.0	14.3	27.4
	2000	13 k	19.9	0.1 k	12.5	4.4	5.0	14.1	32.3
Insur-ance	500	>	1.0 k	0.8 k	56.8	8.0	36.7	71.6	0.1 k
	2000	>	0.3 k	2.5 k	77.2	14.0	33.8	64.7	0.1 k
Mildew	8000	>	>	>	0.5 k	18.2	34.7	0.2 k	0.7 k
	32 000	>	>	>	0.6 k	32.0	0.3 k	0.2 k	0.8 k
Alarm	1000	>	>	>	0.3 k	38.5	47.2	0.3 k	0.7 k
	4000	>	>	>	1.1 k	65.5	47.8	0.3 k	0.7 k

t_{data} represents the runtime of algorithms without prior constraints. t_{LLM} denotes the runtime of algorithms when constrained by LLM-inferred causality. The symbol $>$ signifies a runtime that surpasses 24 hours. Runtimes exceeding 100s are abbreviated for clarity: "0.8k" and "13k" are examples, representing runtimes ranging from 800s to 899s and 13000s to 13999s, respectively.

to 0.9 in the *Child* dataset improves MCMC's performance compared to its outcome with a prior probability of 0.99999 as set in comparison studies.

H. Runtime Evaluation of LLM-Driven Structure Learning

This experiment evaluates the impact of integrating LLM-derived priors on time efficiency. We present a comparison of the average runtime between data-based and LLM-driven structure learning methods in Table XI. The results are individually organized for exact algorithms (DP and A*) and approximate algorithms (Tabu and MCMC). The runtime for NOTEAERS and PC is omitted, as they maintain comparable times between the data-based and LLM-driven versions due to only using edge absence constraints.

For intuitive comparison, we excluded the time spent on prompting LLM, which includes $O(1)$ prompts for λ_p and $O(n)$ prompts for λ_r. We observe that LLM-driven exact search methods are significantly more time-efficient than their data-based counterparts, thanks to the reduction of the hypothesis space by using λ_p as order constraints. In contrast, approximate algorithms experience a runtime increase due to additional checks for path existence in λ_p. Nevertheless, this increase remains manageable as it does not result in an order-of-magnitude difference.

I. Impact of Sample Size

In this experiment, we evaluate the impact of sample size on the performance of our approach. We simulate data with sample sizes ranging from 20 to 5000 and compare the performance of our method with that of the purely data-based algorithm. The F1 score and SHD using MCMC as the backbone algorithm are reported in Fig. 7.

Our results show that our method consistently outperforms the data-driven approach across different sample sizes, with particularly significant improvement at smaller sample sizes. Notably, our LLM-driven causal discovery approach, using only 200 samples, outperforms the data-driven counterpart that uses 1000 or even 5000 samples on many datasets. These findings highlight the potential of our approach to reduce the need for extensive experimental data while achieving superior causal discovery results across various domains.

J. Limitations and Future Direction

Despite the promising results demonstrated by our experiments, there remain certain limitations in our current approach. As discussed in Sections IV-D and IV-E, effective integration of qualitative LLM-derived priors, particularly path existence constraints, into constraint-based and gradient-based structure learning algorithms, is still lacking. This limitation reduces the overall potential of LLM-derived priors to enhance causal discovery performance.

A key direction for future research lies in developing practical strategies to integrate such structural constraints into constraint-based and gradient-based methods. In particular, for gradient-based approaches, there is a significant opportunity to focus on the incorporation of path existence constraints in a way that maintains the advantages of gradient-based optimization. This

would be particularly valuable for employing neural network models to capture complex causal relationships in real-world systems. We believe that advancing these integration techniques will further unlock the power of LLM-derived priors in causal discovery, leading to more robust and practical causal discovery frameworks.

VI. Conclusion

This article proposes a novel approach for integrating LLM-based causal knowledge into data-based structure learning to improve causal discovery. We address the reliability issues in using knowledge-based causality as constraints on data-implied structures by reducing the LLM's influence to reliable causal knowledge. Instead of pairwise causal reasoning that requires both precision and recall, we introduce innovative prompting strategies that focus on one aspect at a time: a focused single-step and decomposed verification strategy to prioritize precision, and an incremental potential cause identification strategy to prioritize recall. By converting these distinct causal insights into a harmonized prior, we offer plug-and-play applications of the prior constraint across mainstream categories of structure learning methods, ensuring broad applicability in practical scenarios.

Experimental results on real-world causal graphs from diverse domains demonstrate the effectiveness of our LLM-driven structure learning in enhancing the performance of data-driven algorithms. These findings highlight the potential of our approach as an autonomous causal discovery tool that facilitates more insightful discoveries across various fields, including biology, healthcare, public policy, and beyond.

References

[1] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[2] R. Opgen-Rhein and K. Strimmer, "From correlation to causation networks: A simple approximate learning algorithm and its application to high-dimensional plant gene expression data," *BMC Syst. Biol.*, vol. 1, pp. 1–10, 2007.

[3] R. Daly, Q. Shen, and S. Aitken, "Learning Bayesian networks: Approaches and issues," *Knowl. Eng. Rev.*, vol. 26, no. 2, pp. 99–157, 2011.

[4] K. Vogel, C. Riggelsen, F. Scherbaum, K. Schröter, H. Kreibich, and B. Merz, "Challenges for Bayesian network learning in a flood damage assessment application," in *Proc. 11th Int. Conf. Struct. Saf. Rel.*, 2013, pp. 16–20.

[5] A. C. Constantinou, Z. Guo, and N. K. Kitson, "The impact of prior knowledge on causal structure learning," *Knowl. Inf. Syst.*, vol. 65, pp. 3385–3434, 2023.

[6] H. Amirkhani, M. Rahmati, P. J. Lucas, and A. Hommersom, "Exploiting experts' knowledge for structure learning of Bayesian networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2154–2170, 2016.

[7] E. Kiciman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," 2023, *arXiv:2305.00050*.

[8] K. Choi, C. Cundy, S. Srivastava, and S. Ermon, "LMPriors: Pre-trained language models as task-specific priors," in *Proc. NeurIPS Found. Models Decis. Mak. Workshop*, 2022. [Online]. Available: https://openreview.net/forum?id=U2MnmJ7Sa4

[9] T. Ban, L. Chen, X. Wang, and H. Chen, "From query tools to causal architects: Harnessing large language models for advanced causal discovery from data," 2023, *arXiv:2306.16902*.

[10] S. Long, A. Piché, V. Zantedeschi, T. Schuster, and A. Drouin, "Causal discovery with language models as imperfect experts," in *Proc. ICML Workshop Structured Probabilistic Inference Generative Model.*, 2023. [Online]. Available: https://openreview.net/forum?id=RXlvYZAE49

[11] A. Vashishtha, A. G. Reddy, A. Kumar, S. Bachu, V. N. Balasubramanian, and A. Sharma, "Causal inference using LLM-guided discovery," 2023, *arXiv:2310.15117*.

[12] A. Abdulaal et al., "Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning," in *Proc. 12th Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=pAoqRlTBtY

[13] C. Liu et al., "Discovery of the hidden world with large language models," 2024, *arXiv:2402.03941*.

[14] K.-H. Cohrs, E. Diaz, V. Sitokonstantinou, G. Varando, and G. Camps-Valls, "Large language models for constrained-based causal discovery," in *Proc. AAAI Workshop "Are Large Lang. Models Simply Causal Parrots?"*, 2023. [Online]. Available: https://openreview.net/forum?id=NEAoZRWHPN

[15] T. Jiralerspong, X. Chen, Y. More, V. Shah, and Y. Bengio, "Efficient causal graph discovery using large language models," 2024, *arXiv:2402.01207*.

[16] P. Li et al., "LLM-enhanced causal discovery in temporal domain from interventional data," 2024, *arXiv:2404.14786*.

[17] Y. Zhang, Y. Zhang, Y. Gan, L. Yao, and C. Wang, "Causal graph discovery with retrieval-augmented generation based large language models," 2024, *arXiv:2402.15301*.

[18] T. Ban, L. Chen, D. Lyu, X. Wang, and H. Chen, "Causal structure learning supervised by large language model," 2023, *arXiv:2311.11689*.

[19] E. Khatibi, M. Abbasian, Z. Yang, I. Azimi, and A. M. Rahmani, "ALCM: Autonomous LLM-augmented causal discovery framework," 2024, *arXiv:2405.01744*.

[20] M. Takayama et al., "Integrating large language models in causal discovery: A statistical causal approach," 2024, *arXiv:2402.01454*.

[21] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social Sci. Comput. Rev.*, vol. 9, no. 1, pp. 62–72, 1991.

[22] D. Zhou et al., "Least-to-most prompting enables complex reasoning in large language models," 2022, *arXiv:2205.10625*.

[23] X. Zhao et al., "Enhancing zero-shot chain-of-thought reasoning in large language models through logic," 2023, *arXiv:2309.13339*.

[24] W. Buntine, "A guide to the literature on learning probabilistic networks from data," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 2, pp. 195–210, Apr. 1996.

[25] X.-W. Chen, G. Anantha, and X. Lin, "Improving Bayesian network structure learning with mutual information-based node ordering in the K2 algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 628–640, May 2008.

[26] H. Chen, B. Liao, J. Luo, W. Zhu, and X. Yang, "Learning a structural causal model for intuition reasoning in conversation," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 3210–3223, Jul. 2024.

[27] L. Zhang, Y. Wu, and X. Wu, "Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 11, pp. 2035–2050, Nov. 2019.

[28] J. Sun et al., "A survey of reasoning with foundation models," 2023, *arXiv:2312.11562*.

[29] C. Zhang et al., "Understanding causality with large language models: Feasibility and opportunities," 2023, *arXiv:2304.05524*.

[30] R. Tu, C. Ma, and C. Zhang, "Causal-discovery performance of ChatGPT in the context of neuropathic pain diagnosis," 2023, *arXiv:2301.13819*.

[31] N. Naik et al., "Applying large language models for causal structure learning in non small cell lung cancer," 2023, *arXiv:2311.07191*.

[32] S. Long et al., "Can large language models build causal graphs?," 2023, *arXiv:2303.05279*.

[33] M. Zečević, M. Willig, D. S. Dhami, and K. Kersting, "Causal parrots: Large language models may talk causality but are not causal," 2023, *arXiv:2308.13067*.

[34] A. Romanou, S. Montariol, D. Paul, L. Laugier, K. Aberer, and A. Bosselut, "CRAB: Assessing the strength of causal relationships between real-world events," 2023, *arXiv:2311.04284*.

[35] S. Lu, I. Bigoulaeva, R. Sachdeva, H. T. Madabushi, and I. Gurevych, "Are emergent abilities in large language models just in-context learning?," 2023, *arXiv:2309.01809*.

[36] Y. Kim, L. Guo, B. Yu, and Y. Li, "Can ChatGPT understand causal language in science claims?," in *Proc. 13th Workshop Comput. Approaches Subjectivity, Sentiment, Soc. Media Anal.*, 2023, pp. 379–389.

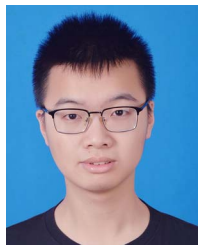
[37] X. Liu et al., "Large language models and causal inference in collaboration: A. comprehensive survey," 2024, *arXiv:2403.09606*.

[38] G. Zhou et al., "Emerging synergies in causality and deep generative models: A survey," 2023, *arXiv:2301.12351*.

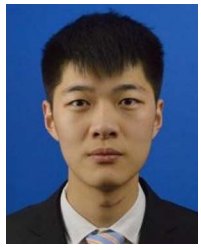
[39] G. Wan, Y. Wu, M. Hu, Z. Chu, and S. Li, "Bridging causal discovery and large language models: A comprehensive survey of integrative approaches and future directions," 2024, *arXiv:2402.11068*.
[40] M. Scutari, C. E. Graafland, and J. M. Gutiérrez, "Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms," *Int. J. Approx. Reasoning*, vol. 115, pp. 235–253, 2019.
[41] W. Lam and A. M. Segre, "A distributed learning algorithm for Bayesian inference networks," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 1, pp. 93–105, Jan./Feb. 2002.
[42] L. Jiang, H. Zhang, and Z. Cai, "A novel bayes model: Hidden naive bayes," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 10, pp. 1361–1371, Oct. 2009.
[43] M. J. Fard, P. Wang, S. Chawla, and C. K. Reddy, "A Bayesian perspective on early stage event prediction in longitudinal data," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3126–3139, Dec. 2016.
[44] N. E. Fenton, M. Neil, and J. G. Caballero, "Using ranked nodes to model qualitative judgments in Bayesian networks," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 10, pp. 1420–1432, Oct. 2007.
[45] H. Wang and D.-Y. Yeung, "Towards Bayesian deep learning: A framework and some existing methods," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3395–3408, Dec. 2016.
[46] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with no tears: Continuous optimization for structure learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9492–9503.
[47] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Mach. Learn.*, vol. 65, pp. 31–78, 2006.
[48] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, no. 10, pp. 2003–2030, 2006.
[49] S. Lee and S. B. Kim, "Parallel simulated annealing with a greedy algorithm for Bayesian network structure learning," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1157–1166, Jun. 2020.
[50] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.
[51] W. Buntine, "Theory refinement on Bayesian networks," in *Uncertainty Proceedings*. Amsterdam, The Netherlands: Elsevier, 1991, pp. 52–60.
[52] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.
[53] D. M. Chickering, "A transformational characterization of equivalent Bayesian network structures," in *Proc. 11th Conf. Uncertainty Artif. Intell.*, 1995, pp. 87–98.
[54] J. A. Gámez, J. L. Mateo, and J. M. Puerta, "Learning Bayesian networks by hill climbing: Efficient methods based on progressive restriction of the neighborhood," *Data Mining Knowl. Discov.*, vol. 22, pp. 106–148, 2011.
[55] A. Li and P. Beek, "Bayesian network structure learning with side constraints," in *Proc. Int. Conf. Probabilistic Graphical Models*, 2018, pp. 225–236.
[56] C. Yuan, B. Malone, and X. Wu, "Learning optimal Bayesian networks using A* search," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 2186–2191.
[57] M. Teyssier and D. Koller, "Ordering-based search: A simple and effective algorithm for learning Bayesian networks," in *Proc. 21st Conf. Uncertainty Artif. Intell.*, 2005, pp. 584–590.
[58] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*. Cambridge, MA, USA: MIT Press, 2000.
[59] N. K. Kitson, A. C. Constantinou, Z. Guo, Y. Liu, and K. Chobtham, "A survey of Bayesian network structure learning," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8721–8814, 2023.
[60] Y. Yu, J. Chen, T. Gao, and M. Yu, "DAG-GNN: DAG structure learning with graph neural networks," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 7154–7163.
[61] K. Bello, B. Aragam, and P. Ravikumar, "DAGMA: Learning DAGs via m-matrices and a log-determinant acyclicity characterization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 8226–8239.
[62] A. Nemirovsky, "Optimization II. Numerical methods for nonlinear continuous optimization," Israel Institute of Technology, 1999.
[63] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 22199–22213.
[64] C. Lee and P. van Beek, "Metaheuristics for score-and-search Bayesian network structure learning," in *Proc. Adv. Artif. Intell.: 30th Can. Conf. Artif. Intell.*, Springer, 2017, pp. 129–141.
[65] R. T. O'Donnell, L. Allison, and K. B. Korb, "Learning hybrid Bayesian networks by MML," in *Proc. 19th Australian Joint Conf. Artif. Intell.*, 2006, pp. 192–203.
[66] T. Silander and P. Myllymäki, "A simple approach for finding the globally optimal Bayesian network structure," in *Proc. Conf. Uncertainty Artif. Intell.*, 2006, pp. 445–452.
[67] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing, "Learning sparse nonparametric dags," in *Proc. Int. Conf. Artif. Intell. Statist.*, PMLR, 2020, pp. 3414–3425.
[68] P. Alquier and G. Biau, "Sparse single-index model," *J. Mach. Learn. Res.*, vol. 14, pp. 243–280, 2013.
[69] Anthropic, *Model Card and Evaluations for Claude Models*, San Francisco, CA, USA: Anthropic, 2023.
[70] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

Taiyu Ban received the BSc degree in computer science and technology from the University of Science and Technology of China, Hefei, China, in 2020. He is currently working toward the PhD degree in computer science and technology with the School of Computer Science and Technology. His current research interests include causal discovery and causal-based machine learning.

Lyuzhou Chen received the BSc degree in mathematics and applied mathematics from the University of Science and Technology of China, Hefei, China, in 2020. He is currently working toward the PhD degree in computer science and technology with the School of Artificial Intelligence and Data Science. His current research interests include causal discovery and causal-based machine learning.



Derui Lyu received the BSc degree in intelligence science and technology from the School of Artificial Intelligence, Xidian University, Xi'an, China, in 2021. He is currently working toward the PhD degree in computer science and technology with the School of Computer Science and Technology, University of Science and Technology of China (USTC), Hefei, China. His current research interests include machine learning and knowledge engineering.



Xiangyu Wang received the BSc degree from the Donghua University, Shanghai, China, and the PhD degree in data science from the University of Science and Technology of China (USTC). He is currently an associate researcher with the School of Computer Science and Technology, USTC. His current research interests include causal discovery and causal-based machine learning.

Qinrui Zhu received the BSc degree in computer science and technology from the University of Science and Technology of China, Hefei, China, in 2020. He is currently working toward the PhD degree in computer science and technology with the School of Computer Science and Technology, and Laboratory for Big Data and Decision. His research interests include large language models.

Huanhuan Chen (Fellow, IEEE) received the BSc degree from the University of Science and Technology of China (USTC), Hefei, China, and the PhD degree in computer science from the University of Birmingham, Birmingham, U.K. He is currently a full professor with the School of Computer Science and Technology, USTC. His current research interests include neural networks, Bayesian inference, and evolutionary computation. He received the 2015 International Neural Network Society Young Investigator Award, the 2012 IEEE Computational Intelligence Society Outstanding PhD Dissertation Award, *IEEE Transactions on Neural Networks* Outstanding Paper Award (bestowed in 2011 and only one paper in 2009), and the 2009 British Computer Society Distinguished Dissertations Award. He is now associate editor of *IEEE Transactions on Neural Networks and Learning Systems*, and *IEEE Transactions on Emerging Topics in Computational Intelligence*.