Causal Structure Learning Supervised by Large Language Model

Taiyu Ban Lyuzhou Chen Derui Lyu Xiangyu Wang* Huanhuan Chen*
School of Computer Science and Technology, University of Science and Technology of China
{banty,clz31415,drlv}@mail.ustc.edu.cn {sa312,hchen}@ustc.edu.cn

***Abstract*—Causal discovery from observational data is pivotal for deciphering complex relationships. Causal Structure Learning (CSL), which focuses on deriving causal Directed Acyclic Graphs (DAGs) from data, faces challenges due to vast DAG spaces and data sparsity. The integration of Large Language Models (LLMs), recognized for their causal reasoning capabilities, offers a promising direction to enhance CSL by infusing it with knowledge-based causal inferences. However, existing approaches utilizing LLMs for CSL have encountered issues, including unreliable constraints from imperfect LLM inferences and the computational intensity of full pairwise variable analyses. In response, we introduce the Iterative LLM Supervised CSL (ILS-CSL) framework. ILS-CSL innovatively integrates LLM-based causal inference with CSL in an iterative process, refining the causal DAG using feedback from LLMs. This method not only utilizes LLM resources more efficiently but also generates more robust and high-quality structural constraints compared to previous methodologies. Our comprehensive evaluation across eight real-world datasets demonstrates ILS-CSL's superior performance, setting a new standard in CSL efficacy and showcasing its potential to significantly advance the field of causal discovery.**

I. Introduction

Causal discovery from the observed data is pivotal in understanding intricate relationships across various domains. Central to this endeavor is Causal Structure Learning (CSL), aiming to construct a causal Directed Acyclic Graph (DAG)[1] from observed data [1]. We adopt causal Bayesian Networks (BNs) as the causal graphical model, renowned for effectively modeling intricate real-world variable relationships [2].

The recovery of high-quality causal BNs faces significant challenges. Firstly, there is the issue of the super-exponential increase in the DAG space as the number of variables grows [3], [4]. Additionally, real-world data is typically sparse and insufficient for accurately representing the true probability distributions [5]. Furthermore, the orientation of edges in a BN cannot be fully deduced from the observed data alone due to the presence of equivalent DAGs [6]. In summary, CSL, when reliant solely on observed data, encounters both practical and theoretical limitations.

Given these inherent limitations, the integration of prior knowledge to constrain specific structures becomes important for reliable causal discovery [7], [8]. While promising, this approach has been limited by the high costs and time associated with expert input [9]. However, the advent of Large Language Models (LLMs) has ushered in a new frontier. Recent studies have underscored the capabilities of LLMs in causal reasoning, positioning them as a valuable and readily accessible resource for knowledge-based causal inference [10], [11], [12].

Kıcıman *et al.* have shown that Large Language Models (LLMs) are effective in determining causality direction between pairs of variables, outperforming even human analysis in this respect [10]. However, other studies highlight LLMs' limitations in constructing causal DAGs from sets of variables, not satisfying even in small-scale contexts [13], [14]. This difficulty mainly stems from the inherent complexity in inferring detailed causal mechanisms, such as establishing the relative directness of causes for an effect, a task that often exceeds simple knowledge-based inference.

In response to these challenges, recent studies have begun integrating LLM-derived causal knowledge with data analysis to enhance causal discovery. For example, Ban *et al.* [15] utilize LLMs to discern the presence of causal links among variables, subsequently applying ancestral constraints to structure learning [7]. This approach yields improvements in learning causal structures from data for smaller-scale problems, but it encounters difficulties with larger datasets due to inaccuracies in the LLM-derived constraints, as evidenced in Table I. As an alternative, Vashishtha *et al.* [16] employ a detailed, pair-based prompting strategy with a voting system to determine reliable prior knowledge. Regretably, the authors fail to show the effectiveness on the larger-scale datasets, likely limited by the complexity and computational demands of the prompt process, which requires $\binom{N}{2}$ LLM inferences with N denoting the variable count.

In response to the challenges, we introduce a simple but effective strategy, named iterative LLM supervised CSL framework (ILS-CSL). Contrasting with prior methodologies that deploy LLMs and CSL separately, ILS-CSL uniquely focuses LLMs on verifying direct causal relationships already suggested by the data. Specifically, ILS-CSL employs LLMs to validate the accuracy of edges in the learned causal DAG, with an iterative process fine-tuning CSL based on LLM feedback. The iteration concludes when the LLM-based inferences and data-driven CSL align within the established causal structure. This innovative integration of LLMs into the CSL process

* These authors are corresponding authors.

[1]In a causal DAG, each edge represents a direct causal link between its nodes.

TABLE I: SHD↓ and constraint quality of the ancestral constraint-based CSL driven by GPT-4, reported in the work [15].

Dataset	Cancer 5 nodes		Asia 8 nodes		Child 20 nodes		Insurance 27 nodes		Alarm 37 nodes		Mildew 35 nodes		Water 32 nodes		Barley 48 nodes	
Data size	250	1000	250	1000	500	2000	500	2000	1000	4000	8000	32000	1000	4000	2000	8000
MINOBSx	3.0	1.8	4.2	2.5	9.5	5.3	25.7	**15.0**	9.5	6.5	**22.8**	**21.0**	**62.3**	**53.7**	**47.0**	**33.7**
+GPT-4	**0.5**	**0.0**	2.2	0.3	10.5	7.8	**24.5**	16.2	12.3	8.8	40.5	21.5	66.7	55.7	52.0	54.5
CaMML	2.0	2.5	3.5	2.2	6.0	1.0	34.3	31.7	11.0	8.2	48.2	62.2	59.0	53.2	81.5	81.2
+GPT-4	2.0	1.3	**0.2**	**0.0**	**4.7**	**1.0**	27.0	22.2	**6.0**	**3.0**	49.2	60.0	58.7	48.3	82.2	82.3
T / F	5 / 0		9 / 0		8 / 2		10 / 0		20 / 1		9 / 6		5 / 3		17 / 7	

The bold SHD is the best performance in each dataset. The cell highlighted in gray indicates a degraded performance by integrating LLM-derived causal knowledge. The row 'T / F' represents the number of correct LLM-derived structrual constraints (T) and that of erroneous ones (F).

offers significant enhancements to the task, as outlined below.

1) **Powerful Structural Constraints:** ILS-CSL transforms the causal inferences made by LLMs into structural constraints explicitly indicating the *edge existence or absence*. The edge-level constraint is more powerful than its path-level counterpart (ancestral constraint) in improving CSL[2], with less risk[3]. Please see Section III-C for further discussions.
2) **Mitigation of Prior Errors:** ILS-CSL markedly diminishes the count of erroneous constraints, all while harnessing identical LLM resources. The reduction is theoretically by a factor of $O(N)$, estimated as $1.8(N-1)$, compared to the full inference on pairwise variables. Please refer to Section V-B for detailed estimation.
3) **Efficient Causal Inference with LLM:** ILS-CSL decreases the number of pairwise variable inferences from $\binom{N}{2}$ to about $O(N)$, as the LLM inference is restricted in the edges of of causal DAG[4]. Such reduction makes the process more manageable and enhances the scalability of the framework.

ILS-CSL has shown consistent improvement in data-driven CSL across all scales of the dataset used in the previous study [15]. It effectively leverages various backbone causal discovery algorithms and demonstrates superior performance, especially as the number of variables increases. These results underscore ILS-CSL's significant potential for facilitating complex causal discovery tasks in real-world scenarios.

II. Related Work

This section discusses the emerging interest in the use of Large Language Models' (LLMs) common sense for understanding causal knowledge. It particularly focuses on the ways this knowledge is being harnessed in causal discovery.

[2]When an ancestral constraint is correctly identified, CSL might still recover a path that includes erroneous edges. In contrast, specifying the existence of an edge directly ensures accuracy, as it cannot be misinterpreted.

[3]An incorrect ancestral constraint inevitably introduces at least one erroneous edge.

[4]Given that the causal DAG is usually sparse, the number of edges $|E|$ is typically estimated as $O(N)$.

A. LLM-based Causal Discovery

Recent advancements in LLM-based causal discovery primarily focus on assessing the inherent capabilities of LLMs [17], [18]. Long *et al.* [14] have tested LLMs' ability to generate simple causal structures, typically with sets of 3-4 variables. In a specialized domain, a study [13] investigates LLMs' effectiveness in discerning causal relationships within medical pain diagnosis, though the findings were somewhat inconclusive.

Kıcıman *et al.* [10] have made strides in optimizing LLM performance for causal analysis by developing more refined prompting techniques. Their work assesses LLMs across a range of causal tasks, revealing notable performance in pairwise causal discovery [19] and counterfactual inference [20], even outperforming human analysis in certain aspects. Additionally, they have enhanced LLMs' capacity to identify causal structures in datasets concerning medical pain diagnosis. However, despite these advancements, a significant gap persists between the quality of causal DAGs generated by LLMs and those derived from data-based algorithms. These findings highlight the potential of LLM-based causal knowledge, yet they also underscore the importance of integrating data in uncovering genuine causal mechanisms.

B. Integration of LLM in Data-based Causal Discovery

A recent work first introduces LLM in causal discovery from data [15]. Recognizing LLMs' limitations in differentiating indirect from direct causality, they applied ancestral constraints based on LLM-generated statements about the existence of causal relationships between variable pairs. The authors prompted the LLM with a complete set of variables, seeking the most confident causal assertions. However, when presented with numerous variables, LLM struggles to provide results that align with causal structures. This complexity leads to a decrease in the accuracy of causal statements as the number of variables increases, as demonstrated in Table **??**. Moreover, we observe that the LLM also fails to make comprehensive causal analyses in larger scale datasets as would be possible with individual prompts for each pair of variables.

Motivated by this work, Vashishtha *et al.* [16] adopted a more targeted method. They individually prompted the LLM

for causal relationships between each variable pair and implemented a voting strategy to deduce ordering constraints. These constraints, although weaker than ancestral constraints (see Section III-C for illustrations), offer more precise structural guidance for causal discovery. Their methodology demonstrates notable improvements across seven real-world datasets. However, the largest dataset examined containes only 23 nodes, leaving the approach's effectiveness in more complex scenarios untested.

III. Preliminaries

We begin by introducing the task of causal structure learning (CSL) on causal Bayesian Networks (BNs) and subsequently discuss the integration of structural constraints.

A. Causal Bayesian Network

A Bayesian Network (BN) is a probabilistic graphical model that uses a Directed Acyclic Graph (DAG) to represent conditional dependencies among a set of variables, thus defining their joint probability distribution. For a set of variables $X = \{X_1, X_2, ..., X_n\}$ in a BN $\mathcal{G}$, the joint probability distribution is given by:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Pa}_i^{\mathcal{G}})$$

$\mathbf{Pa}_i^{\mathcal{G}}$ denotes the parent nodes of X_i in the DAG. It's important to note that an edge in a BN does not inherently imply a causal relationship [1]. A BN can be modified in structures and parameters, and still accurately model the original data distribution, but this does not entail the creation of new causal relationships.

A causal BN, in contrast, not only models the data distribution but also conforms to the principles of causality [21]. In the context of cause-effect relationship, intervening on the causes should render the effect independent of other factors. This introduces additional requirements for representing causality in a BN. In a causal BN, intervening on any subset of variables $X_I \subseteq X$, denoted as $do(X_I = x)$, results in a modified probability distribution $P_I(X)$. This is computed by severing the edges from each variable in X_I to its parents and fixing their values as per the intervention:

$$P_I(X) = \prod_{X_i \notin X_I} P(X_i \mid \mathbf{Pa}_i^{\mathcal{G}}) \quad \text{for all } X \text{ consistent with } x$$

This aspect of causal BNs allows for the modeling of interventions and causal inferences, distinguishing them from standard BNs.

B. Learning Causal BNs

This part introduces the task of two mainstream solutions of learning causal BNs, constraint- and score-based methods. Formally, let $\mathbf{D} \in \mathbb{N}^{m \times n}$ represent the observational data, where m denotes the number of observed samples and n represents the number of observed variables, denoted as $X = \{X_1, X_2, \ldots, X_n\}$. Each X_i in $\mathbf{D}$ takes discrete integer values in the range $[0, C_i)$. Given $\mathbf{D}$, the goal is to determine the causal DAG $\mathcal{G} = (X, E(\mathcal{G}))$, where $E(\mathcal{G})$ denotes the set of directed causal edges among the variables in X. The formal definitions are present as follows:

$$E(\mathcal{G}) \leftarrow \{X_i - X_j \mid X_i \not\perp\!\!\!\perp X_j \mid Y,\ \forall Y \subseteq X \setminus \{X_i, X_j\}\} \tag{1}$$

$$\max_{\mathcal{G}} \sigma(\mathcal{G}; \mathbf{D}) = \sum_{i=1}^{n} \mathcal{L}_\sigma(X_i \mid \mathbf{Pa}_i^{\mathcal{G}}; \mathbf{D}) \text{ s.t. } \mathcal{G} \in \text{DAG} \tag{2}$$

Equations (1) and (2) define the CSL task of constraint- and score-based methods, repectively. Constraint-based methods first determine the skeleton of the graph using undirected edges, $X_i - X_j$, based on conditional independence tests. Subsequently, they orient some of these edges based on V-structure detection and DAG constraints. [22], [23]. Score-based methods employ a scoring function, σ, to evaluate how well a given causal DAG $\mathcal{G}$ represents the observed data $\mathbf{D}$. Typically, σ can be decomposed into scores of local structures, $\mathcal{L}_\sigma(X_i \mid \mathbf{Pa}_i^{\mathcal{G}}; \mathbf{D})$, which simplifies the search process [24], [25]. The objective is to optimize these local scores by assigning appropriate parent nodes to each node, ensuring the resulting graph is a DAG. An alternative approach to searching the DAG space is the ordering-based search, which optimizes Equation (2) under a given ordering O, inherently satisfying the DAG constraint [26], [27]. The best-scored DAG of the searched orderings is then selected as the output.

The design of scoring functions is based on the posterior probability of the DAG given the data, which includes a component representing the prior probability of DAG structures. Due to this adaptability in accommodating the prior constraints on structures, the score-based method is chosen as the backbone CSL algorithm in our ILS-CSL framework.

C. Prior Constraints on Structures

Prior structural constraints play a pivotal role in improving the discovery of causal structures. The most prevalent among these constraints include [28]:

- **Edge Existence**: Denoted as $X_i \to X_j$ or, when forbidden, $X_i \nrightarrow X_j$. This constraint dictates that the DAG should (or should not) contain the edge $X_i \to X_j$.
- **Ordering Constraint**: Represented as $X_i \prec X_j$, it mandates that X_i should precede X_j in the variable ordering.
- **Path Existence (Ancestral Constraint)**: Symbolized as $X_i \rightsquigarrow X_j$, it requires the DAG to encompass the path $X_i \rightsquigarrow X_j$.

Given the implication chain $X_i \to X_j \Rightarrow X_i \rightsquigarrow X_j \Rightarrow X_i \prec X_j$, it is clear that the existence of an edge (direct causality) represents the most stringent structural constraint. Correspondingly, its derivation necessitates a thorough examination of potential combinations of causality. Regrettably, as evidenced by the studies [10], [15], [13], LLMs lack the ability to accurately specify direct causality, often confusing it with indirect causality or non-causal correlations. Please refer to Appendix VII-F for empirical estimation.

Algorithm 1 LLM supervised CSL

Require: Observed data, $\mathbf{D}$; Textual descriptions, $\mathbf{T}$
Ensure: Causal DAG, $\mathcal{G}$
1: Initialize the set of structural constraints, $\lambda \leftarrow \{\}$
2: **repeat**
3: $\quad \mathcal{G} \leftarrow \arg\max_{\mathcal{G}} \sigma(\mathcal{G}; \mathbf{D}),$ s.t. $\mathcal{G} \in \text{DAG}, \mathcal{G} \models \lambda$
4: $\quad$ **for** $X_i \to X_j \in E(\mathcal{G})$ **do**
5: $\quad\quad c \leftarrow$ LLM infers causality between X_i and X_j based on $\mathbf{T}$
6: $\quad\quad$ **if** c is $X_i \leftarrow X_j$ **then**
7: $\quad\quad\quad \lambda \leftarrow \lambda \cup \{X_j \to X_i\}$
8: $\quad\quad$ **end if**
9: $\quad\quad$ **if** c is $X_i \nleftrightarrow X_j$ **then**
10: $\quad\quad\quad \lambda \leftarrow \lambda \cup \{X_i \nrightarrow X_j, X_j \nrightarrow X_i\}$
11: $\quad\quad$ **end if**
12: $\quad$ **end for**
13: **until** no new constraints are added
14: **return** $\mathcal{G}$

Regarding the application of these prior constraints, there are two predominant methodologies: hard and soft approaches. The hard approach prioritizes adherence to prior constraints, followed by score optimization [29]. Conversely, the soft approach strikes a balance between honoring prior constraints and the associated score costs [8]. This often involves adjusting the scoring function to $\sigma(\mathcal{G}; \mathbf{D}) + b(\mathcal{G}; \lambda)$, where a prior probability P_λ is assigned to structural constraints λ. A constraint is only accepted if the bonus score, b, compensates for the penalty in the DAG-data consistency score, σ.

We implement both hard and soft approaches to incorporate structural constraints in this paper.

IV. Iterative LLM Supervised Causal Structure learning

Given the observed data, $\mathbf{D}$, and the descriptive texts on the investigated field and variables, $\mathbf{T}$, the LLM supervised causal structure learning is presented in Algorithm 1.

Initially, a causal DAG $\mathcal{G}$ is learned from $\mathbf{D}$ with modular scoring function $\sigma, \mathcal{L}_\sigma$ (see Equation (2) for definition), and search method $\mathcal{M}$. Subsequently, we explicate the details on LLM supervision and how to constrain CSL accordingly.

A. LLM Supervision

For each directed edge $X_i \to X_j \in E(\mathcal{G})$, we prompt the used LLM to verify the causal statement that X_i `causes` X_j (Line 5 in Algorithm 1). The prompt design for causal inference is inspired by the work [10], which employs choice-based queries to determine the orientation of pairwise variables with known causal relationships. On this basis, we incorporate field-specific descriptions to provide context and introduce additional choices to accommodate uncertainties in causal existence and intricate causal mechanisms. For a given edge $X_i \to X_j$ and associated textual descriptions $\mathbf{T} = \{t_f, t_i, t_j\}$, the LLM is prompted as:

```
You are an expert on t_f. There are two
factors:X_i : t_i, X_j : t_j.
Which cause-and-effect relationship is more likely
for following causal statements for V1 and V2?
A.changing V1 causes a change in V2.
B.changing V2 causes a change in V1.
C.changes in V1 and in V2 are not correlated.
D.uncertain.
Provide your final answer within the tags
<Answer>A/B/C/D</Answer>.
Analyze the statement:X_i X_j.
```

t_f describes the investigated field, and t_i, t_j describes X_i, X_j, respectively. From the LLM's response to this prompt, we can obtain one of the answers: A, B, C, or D.

To specify constraints λ (Lines 6-11 in Algorithm 1), if the answer is B (reversed), we specify the existence of $X_j \to X_i$. If C (no causality), then we specify $X_i \nleftrightarrow X_j$ to forbid the existence of edge. If D (uncertain) or A (correct), we do not specify constraints. This is because specifying the existence of an edge already discovered from data does not often enhance the CSL and can inadvertently lead to errors. For example, if the true structure is $X_i \rightsquigarrow X_j$ but not directly, $X_i \nrightarrow X_j$, LLM easily infers that X_i `causes` X_j due to its shortness in distinguishing indirect causality for the direct. If we specify $X_i \to X_j$, an erroneous edge is introduced.

B. Prior constraint-based CSL

With the structural constraints λ obtained from LLM supervision, we integrate them into the next iteration of CSL process (Line 3 in Algorithm 1), with either hard or soft approach. The process terminates if no new constraint is specified.

a) Hard approach: Firstly, the edge existence and forbidden constraints are used to specify the set of legal candidate parents, $C(i)$, and the set of variables always included in the parents, $K(i)$, of each variable X_i.

$$\begin{aligned} C(i) &= X \setminus \{X_j \mid X_j \nrightarrow X_i \in \lambda\} \setminus \{X_i\} \\ K(i) &= \{X_j \mid X_j \to X_i \in \lambda\} \end{aligned} \tag{3}$$

With $K(i), C(i)$, we prune the space of local structures.

$$L(X_i; \lambda) = \{P \mid K(i) \subseteq P \subseteq C(i)\} \tag{4}$$

The pruned space of local structures, $L(\cdot)$, is taken as input for the search method $\mathcal{M}$:

$$\begin{aligned} &\mathcal{M} : \max_{\mathbf{Pa}_i^{\mathcal{G}}} \sum_i^n \mathcal{L}_\sigma(X_i \mid \mathbf{Pa}_i^{\mathcal{G}}; \mathbf{D}) \\ &\text{s.t. } \mathcal{G} \in \text{DAG}, \mathbf{Pa}_i^{\mathcal{G}} \in L(X_i; \lambda) \end{aligned} \tag{5}$$

In comparison to the problem form without prior constraints, as presented in Equation (2), the restriction of the candidate parent sets of each node, $\mathbf{Pa}_i^{\mathcal{G}} \in L(X_i; \lambda)$, ensures that the output DAG absolutely satisfies every edge constraint, $\mathcal{G} \models \lambda$.

b) Soft approach: We adapt the scoring function to model the edge constraints as follows:

$$\sigma'(\mathcal{G};\mathcal{D},\lambda) = \sum_{i}^{n} \mathcal{L}_\sigma(X_i \mid \mathbf{Pa}_i^{\mathcal{G}};\mathbf{D}) + \mathcal{L}_b(X_i,\mathbf{Pa}_i^{\mathcal{G}};\lambda) \quad (6)$$

$$\begin{aligned}\mathcal{L}_b(X_i,\mathbf{Pa}_i^{\mathcal{G}};\lambda) = \\ \sum_{X_j \to X_i \in \lambda} \left(\mathbb{I}_{X_j \in \mathbf{Pa}_i^{\mathcal{G}}} \log P_\lambda + \mathbb{I}_{X_j \notin \mathbf{Pa}_i^{\mathcal{G}}} \log(1-P_\lambda)\right) \\ + \sum_{X_j \nrightarrow X_i \in \lambda} \left(\mathbb{I}_{X_j \in \mathbf{Pa}_i^{\mathcal{G}}} \log(1-P_\lambda) + \mathbb{I}_{X_j \notin \mathbf{Pa}_i^{\mathcal{G}}} \log P_\lambda\right)\end{aligned} \quad (7)$$

This formulation is grounded in the decomposability of edge constraints. A detailed derivation can be found in Section VI-A. $\mathbb{I}_{\texttt{condition}}$ is the indicator function, which takes the value 1 if the `condition` is true and 0 otherwise. P_λ is the prior confidence, a hyper-parameter. Then search method M optimizes the modified score:

$$\mathcal{M}: \max_{\mathcal{G}} \sum_{i}^{n} \mathcal{L}_\sigma(X_i \mid \mathbf{Pa}_i^{\mathcal{G}};\mathbf{D}) + \mathcal{L}_b(X_i,\mathbf{Pa}_i^{\mathcal{G}};\lambda), \text{ s.t. } \mathcal{G} \in \text{DAG} \quad (8)$$

The bonus score, $\mathcal{L}_b$, favors DAGs that align more closely with the structural constraints. Note that a constraint will not be satisfied if it excessively penalizes the score $\mathcal{L}_\sigma$.

To sum up, while the hard approach derives greater benefits from accurate constraints (at the risk of being more sensitive to errors), the soft approach might not always adhere to all correct constraints but offers a degree of resilience against potential inaccuracies.

V. Analysis on Important Concerns

This section provides in-depth analysis on 1) how does the supervision on the existing skeleton help discovery of missing edges, and 2) the extent to which the prior error is reduced by restricting the LLM inference on the learned causal DAG.

A. Analysis of missing edge discovery

A natural question arises when considering the orientation of the learned skeleton or the prohibition of certain edges: Do these constraints aid in uncovering missing edges? We delve into this question, providing an illustrative analysis for score-based CSL algorithms.

For the sake of discussion, let's assume the ordering of the variables, denoted as O, is given. The ordering naturally satisfies the DAG constraint. Consequently, the score-based search simplifies to a series of independent optimization problems:

$$\max_{\mathbf{Pa}_i^{\mathcal{G}}:\mathbf{Pa}_i^{\mathcal{G}} \subseteq \{X \mid X \prec X_i \text{ in } O\}} \sigma(X_i \mid \mathbf{Pa}_i^{\mathcal{G}}),\ \forall i \in \{1,2,...,N\}$$

where $X \prec X_i$ in O means that node X precedes X_i in the given order O. Given an edge $X_j \to X_i$, forbidding its existence removes X_j from the candidate parent set of X_i. This leads us to the following conclusion:

Lemma 1. *Consider a node X_i in a Bayesian network and its candidate parent variable set C. If X_{opt} represents the optimal parent set of X_i determined by a score-based causal structure learning method, and if a node X_j is removed from C where $X_j \in X_{opt}$, then the newly determined optimal parent set for X_i does not necessarily remain a subset of X_{opt}.*

Lemma 1 is interpreted as that constraining on existing edges can potentially unveil new edges. It's crucial to note that this new edge is distinct from the original skeleton that adheres to O, since a node can not be the parent of its candidate parent node given an ordering.

Viewing this from the lens of knowledge-based causality, constraints derived from known causal relations can enhance the discovery of unknown causal mechanisms within data. This highlights the invaluable role of prior knowledge in advancing causal discovery in uncharted fields.

B. Estimation of Prior Error Counts

Our objective is to estimate and juxtapose the number of erroneous constraints in our framework against that stemming from a full inference on all pairwise variables, an intuitive strategy in the existing methods [15], [16].

We commence by defining four error types and one correctness type that might arise during LLM-based causality inference, along with their respective probabilities:

1) Extra Causality (p_e): Given a causal statement (X_1, X_2), if the true causal DAG neither contains the path $X_1 \rightsquigarrow X_2$ nor $X_2 \rightsquigarrow X_1$, it's an instance of extra causality.
2) Reversed Causality (p_r): Given a causal statement (X_1, X_2), if the true causal DAG contains the path $X_2 \rightsquigarrow X_1$, it's an instance of reversed causality.
3) Reversed Direct Causality (p_r^d): Given a causal statement (X_1, X_2), if the true causal DAG has an edge $X_2 \to X_1$, it's an instance of extra causality.
4) Missing Direct Causality (p_m^d): If an edge $X_1 \to X_2$ or $X_2 \to X_1$ exist in the true causal DAG, but X_1 and X_2 are inferred to have no causal relationship, it's a instance of missing direct causality.
5) Correct Existing Causality (p_c): Given a causal statement (X_1, X_2), if the path $X_1 \rightsquigarrow X_2$ exists in the true causal DAG, it's a instance of correct existing causality.

We assume that the presence of these errors is independent of any specific properties of the pairwise nodes other than its structural relationship. Suppose that the causal DAG comprises N nodes, and the number of pairwise nodes devoid of paths is $\gamma_1\binom{N}{2}$, and the learned causal DAG contains $\gamma_2 N$ edges with a rate of correct edges z_1, reversed edges z_2 and extra edges z_3.

The number of prior errors derived from full inference consists of two parts: the extra causality, $p_e\gamma_1\binom{N}{2}$, and the reversed causality, $p_r(1-\gamma_1)\binom{N}{2}$. Note that the missing causality will not harm the CSL since it does not produce any structural constraints in this context. Then the total number of erroneous constraints is estimated as:

$$E_{\text{full}} = (p_e\gamma_1 + p_r(1-\gamma_1))\binom{N}{2} \quad (9)$$

As for the prior errors within our framework, we consider the output DAG of CSL algorithms. The erroneous constraints

on the correctly discovered edges consist of the reversed and missing direct causality: $(p_r^d + p_m^d)z_1\gamma_2 N$; The erroneous constraints derived from inferring causality on erroneous edges consist of 1) missing direct causality on reversed edges, $p_m^d z_2 \gamma_2 N$, and 2) extra inferred direct causality on extra edges no more than $\left(p_r + p_c P_{R|E}\right) z_3 \gamma_2 N$, where $P_{R|E}$ is the probability where for an extra edge $X_1 \to X_2$ in the learned DAG, a reversed path $X_2 \rightsquigarrow X_1$ exists in the ground truth. Gathering all these, we derive the number prior errors:

$$E_{\text{ours}} \leq \left((p_r^d + p_m^d)z_1 + p_m^d z_2 + (p_r + p_c P_{R|E})z_3\right)\gamma_2 N \tag{10}$$

We random sample pairwise variables on the eight used real-world datasets and prompt GPT-4 to estimate LLM-related parameters p. For estimation of CSL-related ones $\lambda, r, P_{R|E}$, we use outputs of the MINOBSx algorithm, see Appendix VI-B for details. The results are present as:

$$\begin{aligned} &p_e \approx 0.56, p_r \approx 0.15, p_r^d \approx 0.03, p_m^d \approx 0.05 \\ &p_c \approx 0.75, \gamma_1 \approx 0.51, \gamma_2 \approx 1.09, z_1 \approx 0.88 \\ &z_2 \approx 0.05, z_3 \approx 0.07, P_{R|E} \approx 0.05 \end{aligned} \tag{11}$$

And then we have:

$$E_{\text{ours}} \approx 0.10N, E_{\text{full}} \approx 0.36\binom{N}{2}, \frac{E_{\text{ours}}}{E_{\text{full}}} \approx \frac{1}{1.8(N-1)} \tag{12}$$

This indicates that, relative to full pairwise variable inference, ILS-CSL significantly reduces the number of erroneous constraints resulting from imperfect LLM inferences by approximately a factor of $1.8(N-1)$. This reduction is particularly impactful when dealing with larger sets of variables.

VI. Supplementary Illustrations

A. Derivation of Prior-based Scoring

In this section, we derive the prior-based scoring function, as presented in Equations (6) and (7), for the DAG $\mathcal{G}(X, E(\mathcal{G}))$. The prior constraints are denoted as λ : <$\mathbf{R}, \mathbf{\Pi}$>. The set $\mathbf{R} = \{r_1, r_2, \cdots, r_m\}$ comprises edge variables on m pairwise variables, where $r_i \in \{\to, \nrightarrow\}$. $\mathbf{\Pi} = \Pi_{i=1}^m P(r_i)$ is the associated probability distribution.

Beginning with the derivation of the scoring function without prior constraints, let $\mathbf{D}$ be a complete multinomial observed data over variables X. Utilizing the Bayesian Theorem, the probability of a network $\mathcal{G}$ over X is expressed as:

$$P(\mathcal{G}|\mathbf{D}) \propto P(\mathbf{D}|\mathcal{G}) \cdot P(\mathcal{G})$$

Given that $P(\mathbf{D})$ remains consistent across all DAGs, the score of a network is typically the logarithm of $P(\mathcal{G}|\mathbf{D})$, resulting in $Sc(\mathcal{G}|\mathbf{D}) = Sc(\mathbf{D}|\mathcal{G}) + Sc(\mathcal{G})$. Bayesian scoring methods, such as K2 [30] and BDe, BDeu [24], aim to approximate the log-likelihood based on various assumptions. When priors are uniform, $Sc(\mathcal{G})$ can be disregarded during maximization. However, with the introduction of prior structural constraints, denoted as λ, this term gains significance.

Let's define C as a configuration, representing a joint instantiation of values to edge variables $\mathbf{R} = \{r_1, r_2, ..., r_m\}$. The probability for this configuration is $J_C = P(\mathbf{R} = C|\mathbf{\Pi})$. For a specific DAG $\mathcal{G}$, its configuration is represented as $C_\mathcal{G}$. Thus, we can express:

$$P(\mathcal{G} \mid \mathbf{D}, \lambda) = \frac{P(\mathbf{D} \mid \mathcal{G}) \cdot P(\mathcal{G} \mid J)}{P(\mathbf{D} \mid J)} \tag{13}$$

The above equation is derived from the understanding that, given the graph $\mathcal{G}$, the data $\mathbf{D}$ is independent of J. This is because J offers no supplementary information about the data once the graph structure is known. The term $P(\mathbf{D} \mid J)$ serves as a normalizing constant, consistent across all DAGs. The term $P(\mathbf{D} \mid \mathcal{G})$ corresponds to the scoring function $Sc(\mathbf{D} \mid \mathcal{G})$ in the absence of prior constraints. The scoring function can be expressed as:

$$Sc(\mathcal{G} \mid \mathbf{D}, \lambda) = Sc(\mathbf{D} \mid \mathcal{G}) + Sc(\mathcal{G} \mid J) \tag{14}$$

Here, $Sc(\mathbf{D} \mid \mathcal{G})$ represents the scoring function without prior constraints, denoted as $\sigma(\mathcal{G} \mid \mathbf{D})$. Meanwhile, $Sc(\mathcal{G} \mid J)$ pertains to the bonus score associated with prior constraints. Shifting our focus to the prior factor $P(\mathcal{G} \mid J)$, we have:

$$\begin{aligned} P(\mathcal{G} \mid J) =& P\left(\mathcal{G}, C_\mathcal{G} \mid J\right) = P\left(\mathcal{G} \mid J, C_\mathcal{G}\right) \cdot P\left(C_\mathcal{G} \mid J\right) \\ =& P\left(\mathcal{G} \mid C_\mathcal{G}\right) \cdot J_{C_\mathcal{G}} \end{aligned} \tag{15}$$

The first equation holds since $C_\mathcal{G}$ is inherently a function of $\mathcal{G}$. The term $P\left(\mathcal{G} \mid C_\mathcal{G}\right)$ denotes the likelihood of graph $\mathcal{G}$ when a specific configuration is present. In the absence of any other prior constraints, we assign an identical prior to all graphs sharing the same configuration. Let N_C represent the count of DAGs over nodes $\mathcal{V}$ that have the configuration C. Thus, $P\left(\mathcal{G} \mid C_\mathcal{G}\right) = 1/N_{C_\mathcal{G}}$, leading to:

$$P(\mathcal{G} \mid J) = \frac{J_{C_\mathcal{G}}}{N_{C_\mathcal{G}}} \quad \text{and} \quad Sc(\mathcal{G} \mid J) = \log\left(\frac{J_{C_\mathcal{G}}}{N_{C_\mathcal{G}}}\right) \tag{16}$$

Given that the count of edge variables (or edge constraints) remains consistent across all DAGs, $N_{C_\mathcal{G}}$ is also consistent for all DAGs. Therefore:

$$Sc(\mathcal{G} \mid J) = \log J_{C_\mathcal{G}} = \log P(\mathbf{R} = C_\mathcal{G} \mid \mathbf{\Pi}) = \sum_{r_i \in \mathbf{R}} \log P(r_i) \tag{17}$$

Assuming $P(r_i) = P_\lambda$ when λ indicates the presence of the corresponding edge, and $P(r_i) = 1 - P_\lambda$ when the edge's existence is negated, we deduce:

$$\begin{aligned} &Sc(\mathcal{G} \mid J) = \\ &\sum_{X_j \to X_i \in \lambda} \mathbb{I}_{X_j \to X_i \in E(\mathcal{G})} \log P_\lambda + \mathbb{I}_{X_j \to X_i \notin E(\mathcal{G})} \log(1 - P_\lambda) + \\ &\sum_{X_j \nrightarrow X_i \in \lambda} \mathbb{I}_{X_j \to X_i \in E(\mathcal{G})} \log(1 - P_\lambda) + \mathbb{I}_{X_j \to X_i \notin E(\mathcal{G})} \log P_\lambda \end{aligned} \tag{18}$$

By integrating Equations (14), (18), and (2), we derive the form of the local prior constraint-based scoring function, as depicted in Equations (6) and (7).

TABLE II: Accuracy and reversed ratio of the sampled pairwise variables on eight datasets.

Dataset	Alarm	Asia	Insurance	Mildew	Child	Cancer	Water	Barley
Direct causality (Acc_1 /Rev_1)	1.00 / 0.00	1.00 / 0.00	0.85 / 0.05	0.95 / 0.05	1.00 / 0.00	1.00 / 0.00	0.95 / 0.05	0.70 / 0.05
Indirect causality (Acc_2 /Rev_2)	0.65 / 0.15	1.00 / 0.00	0.95 / 0.05	1.00 / 0.00	0.50 / 0.40	1.00 / 0.00	0.50 / 0.50	0.30 / 0.30
No causality (Acc_3)	0.60	0.80	0.35	0.10	0.50	0.00	0.45	0.50
Qualitative causality(Acc_4 / Rev_4)	0.72 / 0.12	1.00 / 0.00	0.92 / 0.05	0.99 / 0.01	0.70 / 0.24	1.00 / 0.00	0.67 / 0.33	0.36 / 0.26

B. Parameter Estimation in Section V-B

This section presents the details on the estimation of parameters related to the quality of LLM based causal inference, $p_e, p_r, p_r^d, p_m^d, p_c$, structures of the true causal DAGs, γ_1, and structures of the learned causal DAGs, $\gamma_2, z_1, z_2, z_3, P_{R|E}$.

a) Quality of LLM causal inference: We randomly sample three kinds of pairwise variables from the employed eight datasets in experiments:

1) Direct edges: Sampling pairwise variables with direct edge $X_i \to X_j$ in the ground truth.
2) Indirect path: Sampling pairwise variables without direct edge but with a directed path, $X_i \nrightarrow X_j, X_i \rightsquigarrow X_j$.
3) Not connected: Sampling pairwise variables without any path, $X_i \not\rightsquigarrow X_j, X_j \not\rightsquigarrow X_i$.

For each type, we sample 20 pairwise variables form each dataset, if more than 20 pairwise variables satisfying the condition exist in the causal DAG. Or we use all the pairwise variables as samples.

Subsequently, we query GPT-4 the causality between each pairwise variables through the prompt in Section IV. The true answer of Types 1 and 2 is A, and that of Type 3 is C. The accuracy of GPT-4 on different datasets on these samples together with the ratio of reversed inference (B for Types 1 and 2) are reported in Table II.

Direct causality corresponds to direct edges, indirect causality to indirect paths, and no causality corresponds to not connected variables. The accuracy and reversed ratio of LLM inference on them is obtained by experiments. The qualitative causality corresponds the paths (including edges), whose accuracy is estimated by $\text{Acc}_4 = (\text{Acc}_1 \times |E| + \text{Acc}_2 \times |P|)/(|E| + |P|)$, where $|E|$ and $|P|$ represents the number of edges and indirect paths in the true causal DAG.

By weighted sum of the accuracy and reversed ratio, we obtain the estimation of them. Then the probability of the five introduced error that GPT-4 makes are presented as follows:

1) Extra causality: $p_e = 1 - \text{Acc}_3 = 0.56$
2) Reversed causality: $p_r = \text{Rev}_4 = 0.15$
3) Reversed direct causality: $p_r^d = \text{Rev}_1 = 0.03$
4) Missing direct causality: $p_m^d = 1 - \text{Acc}_1 - \text{Rev}_1 = 0.05$
5) Correct existing causality: $p_c = \text{Acc}_4 = 0.75$

We see that the major errors of GPT-4 inference is sourced from the extra causality, which is because some intuitively correlated concepts may not generate real causal relations in an experiment with specific conditions. And that is why we should refer to data for causal analysis. However, GPT-4 is prone to infer correct causality on pairwise variables with direct causality, which is the base of our framework to efficiently improves the quality of learned causal DAGs.

b) Structural parameters: The structural parameters is estimated by the average value of them on the eight datasets. The ones related to the causal structure learning of each dataset is estimated by the average value of them on twelve segments of observed data, using MINOBSx search and BDeu score. See the detailed results in Table III.

TABLE III: The estimated structural paramters on eight datasets.

Dataset	Alarm	Asia	Insurance	Mildew	Child	Cancer	Water	Barley	Avg.
γ_1	0.67	0.36	0.52	0.52	0.66	0.20	0.65	0.52	0.51
γ_2	1.22	1.01	1.44	0.79	1.09	0.55	1.34	1.27	1.09
z_1	0.96	0.88	0.91	0.87	0.98	0.90	0.67	0.84	0.88
z_2	0.02	0.00	0.05	0.08	0.00	0.07	0.12	0.07	0.05
z_3	0.02	0.12	0.04	0.05	0.02	0.03	0.21	0.09	0.07
$P_{R\|E}$	0.02	0.00	0.05	0.08	0.00	0.10	0.12	0.08	0.05

VII. Experiments

We conduct experiments to address the research questions:
RQ1: Can ILS-CSL enhance data-based CSL baselines and outperform the existing LLM-driven CSL method?
RQ2: Across diverse backbone algorithms, can ILS-CSL consistently improve the quality of causal structures? which of the soft and hard constraint is better?
RQ3: Is ILS-CSL resistant to imperfect LLM causal inferences, and capable to derive accurate prior? Why?
RQ4: How does the process, where LLM supervises causal discovery, unfold in detail?

All the datasets, codes, and supplementary results can be accessed in the external repository[5].

A. Datasets and Baselines

To address RQ1, we employ the eight real-world datasets of causal DAGs from the Bayesian Network Repository[6] as used in the comparative study [15]. Dataset specifics are provided in Table IV. For backbone CSL algorithms, we adopt the same MINOBSx (BDeu score) [28] and CaMML (MML score) [31] algorithms, and utilize the same setting of prior probability for CaMML, 0.99999. For supervision on CSL, we utilize GPT-4-WEB[7]. For RQ2, the used baselines comprise a combination of popular scoring functions, namely BIC and BDeu score [24], and search algorithms, including HC [32] and MINOBSx [33].

[5]https://github.com/tyMadara/ILS-CSL
[6]https://www.bnlearn.com/bnrepository/
[7]https://chat.openai.com/

TABLE IV: The used datasets of causal DAGs.

Dataset	Cancer	Asia	Child	Alarm	Insurance	Water	Mildew	Barley
Variables	5	8	20	37	27	32	35	48
Edges	4	8	25	46	52	66	46	84
Parameters	10	18	230	509	1008	10083	540150	114005
Data size	250 / 1000	250 / 1000	500 / 2000	1000 / 4000	500 / 2000	1000 / 4000	8000 / 32000	2000 / 8000

TABLE V: Scaled SHD↓ comparison to data-based and LLM-driven CSL.

Dataset	Cancer		Asia		Child		Insurance	
N	250	1000	250	1000	500	2000	500	2000
MINOBSx	0.75±0.22	0.46±0.29	0.52±0.32	0.31±0.07	0.38±0.08	0.21±0.04	0.46±0.05	0.29±0.02
+sepLLM-hard	0.13 -83%	0.00 -100%	0.27 -48%	0.04 -87%	0.42 +11%	0.31 +48%	0.91 +98%	0.60 +107%
+ILS-CSL-hard	0.50±0.22 -33%	0.29±0.29 -37%	0.42±0.37 -19%	0.15±0.15 -52%	0.25±0.06 -34%	0.07±0.03 -67%	0.42±0.03 -9%	0.28±0.06 -3%
CaMML	0.75±0.00	0.62±0.14	0.58±0.29	0.27±0.05	0.25±0.03	0.09±0.04	0.69±0.04	0.61±0.15
+sepLLM-soft	0.50 -33%	0.33 -47%	0.02 -97%	0.00 -100%	0.19 -24%	0.04 -56%	1.00 +45%	0.82 +34%
+ILS-CSL-soft	0.75±0.00 +0%	0.33±0.20 -47%	0.23±0.09 -60%	0.15±0.18 -44%	0.17±0.05 -32%	0.04±0.00 -56%	0.47±0.04 -32%	0.47±0.11 -23%
Dataset	Alarm		Mildew		Water		Barley	
N	1000	4000	8000	32000	1000	4000	2000	8000
MINOBSx	0.21±0.06	0.14±0.04	0.50±0.02	0.46±0.05	0.77±0.07	0.61±0.04	0.56±0.04	0.40±0.03
+sepLLM-hard	0.27 +29%	0.19 +36%	0.88 +76%	0.47 +2%	1.01 +31%	0.84 +38%	0.62 +11%	0.65 +62%
+ILS-CSL-hard	0.09±0.03 -57%	0.08±0.02 -43%	0.43±0.00 -14%	0.33±0.18 -28%	0.68±0.05 -12%	0.56±0.02 -8%	0.54±0.02 -4%	0.38±0.02 -5%
CaMML	0.24±0.05	0.18±0.06	1.20±0.10	1.30±0.12	0.88±0.08	0.81±0.04	0.96±0.07	0.96±0.10
+sepLLM-soft	0.13 -46%	0.07 -61%	1.07 -11%	1.30 +0%	0.89 +1%	0.73 -10%	0.98 +2%	0.98 +2%
+ILS-CSL-soft	0.08±0.01 -67%	0.06±0.01 -67%	1.01±0.07 -16%	1.26±0.05 -3%	0.70±0.02 -20%	0.63±0.04 -22%	0.90±0.06 -6%	0.83±0.06 -14%

The suffixes '-hard' and '-soft' represent the approach to apply the LLM inferred prior constraints. The performances of sepLLM method are obtained from the work [15].

B. Observed Data and Evaluation Metric

We utilize a collection of observed data sourced from a public repository[8]. This data, generated based on the eight causal DAGs, is provided by Li and Beek [28], and used in the comparative work [15]. The repository offers datasets in two distinct sample sizes for each DAG, as detailed in Table IV. For every sample size, six distinct data segments are available.

To assess the quality of the learned causal structures, we primarily employ the scaled Structural Hamming Distance (SHD) [34]. This metric is defined as the SHD normalized by the total number of edges in the true causal DAG.

TABLE VI: Ranking of methods in Table V.

Data-based CSL		SepLLM		ILS-CSL	
MINOBSx	CaMML	MINOBSx	CaMML	MINOBSx	CaMML
3.6	4.8	4.0	3.9	**1.9**	2.9

C. Comparison Experiments (RQ1)

We compare the performance of MINOBSx (BDeu) and CaMML that are used in the separate LLM prior-driven CSL approach proposed by [15], referred to as sepLLM, and our proposed framework, termed ILS-CSL. This comparison is conducted using all the introduced observed data across eight datasets. The results, presented in terms of scaled SHD (where a lower value is preferable), are detailed in Table V. The difference between scaled SHD of data-based (Δ_{data}) and LLM-driven (Δ_{LLM}) CSL is also reported, by calculating $(\Delta_{\text{LLM}} - \Delta_{\text{data}})/\Delta_{\text{data}}$. The Friedman ranking of the methods and more is reported in Table VI.

Key observations from Table V are presented as follows.

1) ILS-CSL consistently improves the quality of data-based CSL in all cases, with the sole exception observed in the *Cancer* dataset with 250 samples, where it maintains the same performance. In contrast, sepLLM shows consistent improvement only in the *Cancer* and *Child* datasets, while exhibiting partial performance degradation in others. This observation underscores the robust and stable enhancement offered by our ILS-CSL framework.
2) Our framework outperforms sepLLM in datasets with more than 20 variables, albeit showing lesser performance in small-scale datasets, *Cancer* and *Asia*. This trend is attributed to the relatively simple causal mechanisms in these smaller datasets, where LLM effectively infers correct causal relationships between variables (refer to Table II in Appendix VI-B). Despite sepLLM leveraging all existing causality inferred by LLM, its advantage is pronounced only in these two datasets. As the complexity of causal mechanisms increases with the number of variables, the quality of LLM inference diminishes, highlighting the resilience of our framework against imperfect LLM inference.

Table VI demonstrates that ILS-CSL consistently ranks within the top two positions. Notably, within the sepLLM framework, CaMML, which uses soft constraints, outperforms MINOBSx, which relies on hard constraints. However, this trend reverses in the ILS-CSL framework. This shift is at-

[8]https://github.com/andrewli77/MINOBS-anc/tree/master/data/csv

TABLE VII: Scaled SHD↓ enhancement on data-based CSL with different scores, search algorithms and approaches to apply prior constraints, by the proposed framework.

Dataset	Cancer		Asia		Child		Insurance	
N	250	1000	250	1000	500	2000	500	2000
HC-BDeu	0.58±0.13	0.33±0.26	0.56±0.27	0.23±0.17	0.57±0.12	0.49±0.18	0.69±0.06	0.68±0.09
+ILS-CSL-hard	0.50±0.22 -14%	0.29±0.29 -12%	0.46±0.33 -18%	0.15±0.15 -35%	0.24±0.07 -58%	0.10±0.02 -80%	0.45±0.06 -35%	0.34±0.04 -50%
+ILS-CSL-soft	0.50±0.22 -14%	0.29±0.29 -12%	0.44±0.30 -21%	0.15±0.15 -35%	0.26±0.06 -54%	0.11±0.03 -78%	0.50±0.08 -28%	0.35±0.04 -49%
MINOBSx-BDeu	0.75±0.22	0.46±0.29	0.52±0.32	0.31±0.07	0.38±0.08	0.21±0.04	0.46±0.05	0.29±0.02
+ILS-CSL-hard	0.50±0.22 -33%	0.29±0.29 -37%	0.42±0.37 -19%	0.15±0.15 -52%	0.25±0.06 -34%	0.07±0.03 -67%	0.42±0.03 -9%	0.28±0.06 -3%
+ILS-CSL-soft	0.50±0.22 -33%	0.29±0.29 -37%	0.42±0.37 -19%	0.15±0.15 -52%	0.25±0.04 -34%	0.08±0.04 -62%	0.41±0.03 -11%	0.26±0.04 -10%
HC-BIC	0.92±0.29	0.62±0.34	0.48±0.36	0.31±0.29	0.53±0.07	0.38±0.16	0.76±0.05	0.72±0.06
+ILS-CSL-hard	0.92±0.29 +0%	0.42±0.34 -32%	0.33±0.25 -31%	0.19±0.17 -39%	0.26±0.07 -51%	0.07±0.03 -82%	0.60±0.03 -21%	0.41±0.03 -43%
+ILS-CSL-soft	0.92±0.29 +0%	0.42±0.34 -32%	0.35±0.26 -27%	0.21±0.19 -32%	0.27±0.08 -49%	0.07±0.05 -82%	0.62±0.06 -18%	0.42±0.03 -42%
MINOBSx-BIC	1.00±0.25	0.62±0.21	0.46±0.23	0.27±0.05	0.34±0.06	0.18±0.04	0.62±0.05	0.55±0.05
+ILS-CSL-hard	0.92±0.29 -8%	0.38±0.26 -39%	0.42±0.40 -9%	0.12±0.08 -56%	0.24±0.08 -29%	0.06±0.02 -67%	0.55±0.03 -11%	0.39±0.08 -29%
+ILS-CSL-soft	0.92±0.29 -8%	0.38±0.26 -39%	0.35±0.26 -24%	0.15±0.12 -44%	0.25±0.05 -26%	0.06±0.02 -67%	0.55±0.03 -11%	0.41±0.09 -25%

Dataset	Alarm		Mildew		Water		Barley	
N	1000	4000	8000	32000	1000	4000	2000	8000
HC-BDeu	0.65±0.12	0.64±0.09	0.79±0.11	0.99±0.07	0.76±0.07	0.64±0.08	0.80±0.06	0.65±0.06
+ILS-CSL-hard	0.12±0.02 -82%	0.08±0.01 -88%	0.46±0.01 -42%	0.22±0.02 -78%	0.64±0.02 -16%	0.55±0.03 -14%	0.69±0.06 -14%	0.57±0.06 -12%
+ILS-CSL-soft	0.30±0.05 -54%	0.25±0.06 -61%	0.43±0.00 -46%	0.47±0.04 -53%	0.64±0.01 -16%	0.56±0.03 -12%	0.76±0.04 -5%	0.62±0.03 -5%
MINOBSx-BDeu	0.21±0.06	0.14±0.04	0.50±0.02	0.46±0.05	0.77±0.07	0.61±0.04	0.56±0.04	0.40±0.03
+ILS-CSL-hard	0.09±0.03 -57%	0.08±0.02 -43%	0.43±0.00 -14%	0.33±0.18 -28%	0.68±0.05 -12%	0.56±0.02 -8%	0.54±0.02 -4%	0.38±0.02 -5%
+ILS-CSL-soft	0.09±0.02 -57%	0.07±0.01 -50%	0.47±0.01 -6%	0.37±0.02 -20%	0.68±0.04 -12%	0.56±0.02 -8%	0.55±0.03 -2%	0.38±0.02 -5%
HC-BIC	0.68±0.05	0.59±0.10	0.90±0.06	0.91±0.13	0.76±0.04	0.70±0.03	0.87±0.05	0.80±0.08
+ILS-CSL-hard	0.22±0.04 -68%	0.12±0.04 -80%	0.58±0.01 -36%	0.46±0.04 -49%	0.69±0.02 -9%	0.61±0.03 -13%	0.76±0.02 -13%	0.69±0.06 -14%
+ILS-CSL-soft	0.41±0.04 -40%	0.35±0.11 -41%	0.71±0.01 -21%	0.57±0.02 -37%	0.69±0.02 -9%	0.61±0.03 -13%	0.82±0.04 -6%	0.74±0.09 -8%
MINOBSx-BIC	0.32±0.08	0.15±0.04	0.74±0.01	0.73±0.09	0.82±0.03	0.77±0.03	0.79±0.04	0.58±0.03
+ILS-CSL-hard	0.16±0.07 -50%	0.09±0.03 -40%	0.58±0.01 -22%	0.45±0.03 -38%	0.69±0.03 -16%	0.62±0.01 -19%	0.73±0.03 -8%	0.55±0.03 -5%
+ILS-CSL-soft	0.19±0.06 -41%	0.10±0.01 -33%	0.73±0.01 -1%	0.64±0.04 -12%	0.70±0.02 -15%	0.64±0.02 -17%	0.76±0.02 -4%	0.56±0.03 -3%

TABLE VIII: Ranking of methods in Table VII.

BDeu						BIC					
MIN-OBSx	+hard	+soft	HC	+hard	+soft	MIN-OBSx	+hard	+soft	HC	+hard	+soft
7.0	**2.7**	2.8	10.1	3.4	5.3	9.8	4.9	6.2	11.2	6.6	8.0

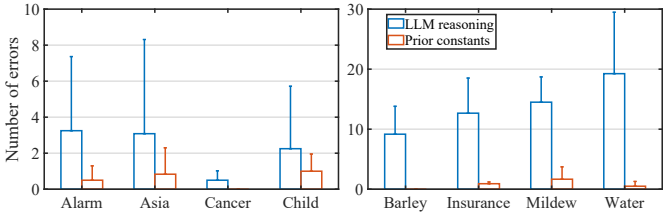


Fig. 1: Erroneous LLM inference and erroneous specified edge constraints of MINOBSx-BDeu+ILS-CSL-hard.

tributed to the ability of soft constraints to filter out some incorrect prior structures that significantly conflict with the data distribution. The prior constraints in sepLLM are not as high-quality as those in LLM-CSL, and the use of ancestral constraints in sepLLM tends to introduce erroneous edges.

D. ILS-CSL With Diverse Backbone Algorithms (RQ2)

We experiment with varying scoring functions, BDeu and BIC scores, and search algorithms, MINOBSx and HC, and compare to corresponding data-based CSL performances. Moreover, we experiment with both hard and soft approaches to apply prior constraints, with the prior probability setting $P_\lambda = 0.99999$ introduced in Equation (7). The results on the utilized observed data of eight datasets are reported in Table VII. The Friedman ranking of the methods is reported in Table VIII. Key observations include:

1) Nearly all scenarios showcase an enhancement, underscoring the impactful role of ILS-CSL in improving CSL performance across diverse datasets and algorithms.
2) ILS-CSL's impact on causal discovery significantly surpasses the limitations imposed by scoring functions and search algorithms. The ranking results demonstrate this clearly, as HC+ILS-CSL exceeds the performance of MINOBSx, even with a less robust baseline. This also holds true across different scoring functions, highlighting ILS-CSL's broad applicability and effectiveness in improving causal discovery outcomes.
3) The hard approach outperforms the soft approach, attributed to the high quality of specified constraints within ILS-CSL. This stands in stark contrast to the findings by [15], where the soft approach fared better due to the lower quality of prior constraints.

E. Errors in LLM Inference and Prior Constraints (RQ3)

This section is dedicated to the evaluation of ILS-CSL's robustness against the inaccuracies in LLM inference. We scrutinize the erroneous causal relationships inferred by LLM on the edges of the learned DAG, along with the incorrect

TABLE IX: The precision along with ratio of different structures of different answers by GPT-4.

Answer	Dataset	Direct edges	Reversed edges	Precision	Indirect paths	Reversed indirect paths	Not reachable	Overall Precision	
								Qualitative	Structural
A	Alarm	0.33	0.02	0.94	0.28	0.00	0.37	0.61	0.33
	Asia	0.44	0.00	1.00	0.50	0.00	0.06	0.94	0.44
	Barley	0.22	0.12	0.65	0.23	0.12	0.31	0.45	0.22
	Cancer	0.36	0.09	0.80	0.36	0.09	0.09	0.73	0.36
	Child	0.46	0.02	0.96	0.26	0.04	0.22	0.72	0.46
	Insurance	0.41	0.05	0.89	0.32	0.06	0.15	0.74	0.41
	Mildew	0.45	0.04	0.92	0.36	0.03	0.11	0.82	0.45
	Water	0.47	0.13	0.78	0.11	0.01	0.28	0.58	0.47
B	Alarm	0.02	0.36	0.95	0.10	0.18	0.34	0.54	0.36
	Asia	0.00	0.50	1.00	0.00	0.36	0.14	0.86	0.50
	Barley	0.02	0.21	0.91	0.08	0.43	0.25	0.64	0.21
	Cancer	0.00	0.60	1.00	0.00	0.00	0.40	0.60	0.60
	Child	0.00	0.45	1.00	0.24	0.12	0.18	0.58	0.45
	Insurance	0.02	0.59	0.97	0.02	0.10	0.27	0.68	0.59
	Mildew	0.01	0.49	0.98	0.00	0.14	0.35	0.64	0.49
	Water	0.03	0.51	0.94	0.29	0.03	0.14	0.54	0.51
C	Alarm	0.00	0.00	-	0.00	0.03	0.97	0.97	1.00
	Asia	0.00	0.00	-	0.00	0.00	1.00	1.00	1.00
	Barley	-	-	-	-	-	-	-	-
	Cancer	-	-	-	-	-	-	-	-
	Child	0.00	0.11	-	0.00	0.11	0.79	0.79	0.89
	Insurance	0.03	0.05	-	0.00	0.10	0.83	0.83	0.93
	Mildew	0.00	0.01	-	0.32	0.36	0.32	0.32	0.99
	Water	0.00	0.04	-	0.30	0.19	0.47	0.47	0.96

prior constraints that stem from them. The results pertaining to each dataset, which includes two unique sizes of observed data related to MINOBSx-BDeu with the hard approach, are illustrated in Figure 1. For a more comprehensive set of results, refer to the external repository.

Our observations highlight a substantial reduction in the errors of specified edge constraints compared to erroneous LLM inference. This reduction stems from the strategy of only imposing constraints on causality that is inconsistent with what has been learned. A more detailed analysis on the superior aspect of ILS-CSL to reduce erroneous constraints is made in the following experiment.

F. Why Resistant to Imperfect LLM Inference (RQ3)

This section elucidates the ability of ILS-CSL to minimize prior errors by limiting LLM supervision to edges. We present the ratio of various real structures corresponding to all pairwise variables inferred by GPT-4. Table IX displays the results for all datasets, highlighting the precision related to ILS-CSL (light red cells) and full inference (light blue cells). It distinguishes between qualitative precision (correct paths) and structural precision (correct edges only).

In the context of the analysis, the outcomes A, B, and C from GPT-4 have specific meanings related to inferred causal relationships between two variables X_1 and X_2:
Outcome A: GPT-4 infers that X_1 causes X_2 ($X_1 \to X_2$).
Outcome B: GPT-4 infers that X_2 causes X_1 ($X_2 \to X_1$).
Outcome C: GPT-4 infers that X_1 and X_2 are not causally related ($X_1 \nleftrightarrow X_2$).

In the table, various columns represent the ratio of different corresponding structures in the ground truth:
Direct Edges: The edge ($X_1 \to X_2$) exists in truth.
Reversed Edges: An reversed edge ($X_2 \to X_1$) exists in truth.
Indirect Paths: A path ($X_1 \rightsquigarrow X_2$) exists, but ($X_1 \nrightarrow X_2$).
Reversed Indirect Paths: ($X_2 \rightsquigarrow X_1$), but ($X_2 \nrightarrow X_1$).
Not Reachable: ($X_1 \not\rightsquigarrow X_2, X_2 \not\rightsquigarrow X_1$).

The precision of LLM on variables that have edges (light red cells of answers A and B) is notably high, significantly exceeding the precision on variables that may not. Analyzing prior errors in ILS-CSL reveals:

1) For GPT-4 outcome C, the corresponding edge forbidden constraints exhibit high precision, generating few erroneous structural constraints. This is attributed to the high confidence in the absence of causal relations inferred based on knowledge, leading to excellent precision on pairwise variables without structural edges, albeit with a lower recall.
2) For GPT-4 outcomes A or B, high precision is observed on learned edges belonging to the true skeleton, producing few erroneous structural constraints. Given known direct causality between pairwise variables, LLM can easily infer the correct causal direction, stemming from the counterintuitive nature of reversed causal statements.
3) Major LLM inference errors stem from outcomes A and B on learned edges outside the true skeleton. However, the impact of these errors on generating incorrect structural constraints is mitigated by the low probability of extra edges occurring in a learned structure ($z_3 \approx 0.07$,

see Table III) and the strategy of specifying a prior constraint only when inconsistent.

In essence, the primary limitation of LLM in causal inference is the confusion between direct causal relationships, indirect causality, and correlations, evidenced by the low overall qualitative and structural precision. This limitation hampers the performance of using LLM-derived existence on causality as ancestral (qualitative precision) or edge constraints (structural precision) seperately.

Contrarily, ILS-CSL effectively minimizes prior errors by leveraging the inherent precision of LLM in inferring non-causal relations and determining causal direction on pairwise variables with direct causality. It smartly circumvents LLM's limitation in discerning the existence of direct causal relationships, which are easily confused with indirect causality or correlations, by restricting the LLM inference into the range of learned structures from data, as analyzed in point 3.

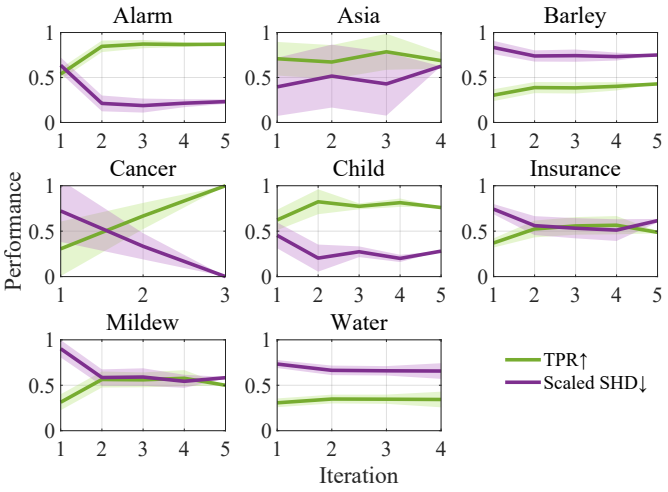


Fig. 2: Trend of TPR↑ (green line) and scaled SHD↓ (purple line) of HC+BIC+ILS-CSL-hard on various datasets.

G. Trend of DAG Quality over Iterations (RQ4)

This section outlines the iterative trends of scaled SHD (aiming for a decrease, denoted as SHD↓) and True Positive Rate (aiming for an increase, denoted as TPR↑) for various backbone algorithms across eight datasets. Each dataset spans two distinct data sizes, resulting in 12 segments of observed data. It's crucial to note the potential for significant derivation due to performance differences across varying data sizes, particularly for smaller-scale datasets like *Cancer* and *Asia*. The results of HC+BIC+ILS-CSL-hard on various datasets are reported in Figure 6, with comprehensive results available in the external repository. Key observations from the iterative trends include:

- **Limited Iteration Numbers:** Most cases require a limited number of iterations. The area near the maximum iteration in each figure is small when exceeding 4, indicating that few out of the 12 cases reach this point. Some cases even have a derivation of zero at the maximum iteration, signifying that only one case attains this maximum value.
- **Quality Improvement Trend:** Generally, as the iteration number increases, the scaled SHD decreases, and the TPR increases. This trend underscores the enhancement in the quality of the learned causal structures as ILS-CSL progresses.
- **Significant Initial Improvement:** The most substantial improvement in the quality of learned causal DAGs occurs in the first round of LLM supervision (from Iteration 1 to 2). Subsequent iterations offer diminished enhancements. This pattern is attributed to the initial presentation of most inconsistent edges with LLM inference in the first iteration. Post the integration of prior constraints, the new structures learned by CSL exhibit far fewer inconsistencies with LLM inference.
- **Potential Quality Degradation:** In certain instances, the quality of the causal DAG diminishes across specific iterations. This decline could stem from the introduction of new erroneous prior constraints in a given iteration or a statistical artifact. The latter scenario arises when two consecutive iterations do not employ the same set of observed data, as some cases conclude in the preceding iteration.

These observations provide a comprehensive insight into the iterative behavior of ILS-CSL, highlighting its effectiveness and areas of caution to ensure consistent enhancement in learned causal structures.

H. Illustrative Example of DAG Evolution (RQ4)

We visualize the learned causal structures in iterations to unfold the details of ILS-CSL. An illustrative example by HC (BDeu) algorithm on *Child* dataset, 2000 samples, with hard constraining approach in ILS-CSL, is reported in Figure 3.

Initially, HC (BDue) learns a causal DAG from pure observed data (Iteration 0), whose edges are supervised by LLM, leading to edge constraints (colored arrows) on inconsistent inferred edge by LLM. The constraints could refine local structures (red arrows) or bring harm due to the erroneous inference (blue arrows). The erroneous edges (dotted arrows) are reduced as the iteration goes. Details of further observations are presented as follows:

- The SHD of the learned causal DAG is greatly reduced from 12 to 3 by employing the ILS-CSL framework, showcasing the significant capability of our framework to enhance the quality of learned causality.
- The first round of LLM-based supervision refines the learned DAG to a much greater extent than the following rounds. This addresses the acceptable efficiency loss of ILS-CSL, which usually does not require many iterations.
- There are 7 correct constraints (red arrow) and 2 erroneous ones (blue arrow) in total. The number of directly corrected edges by these priors is $7 - 2 = 5$, while the reduced SHD is 8, meaning that 3 edges that are distinct from those in constraints are corrected without any prior

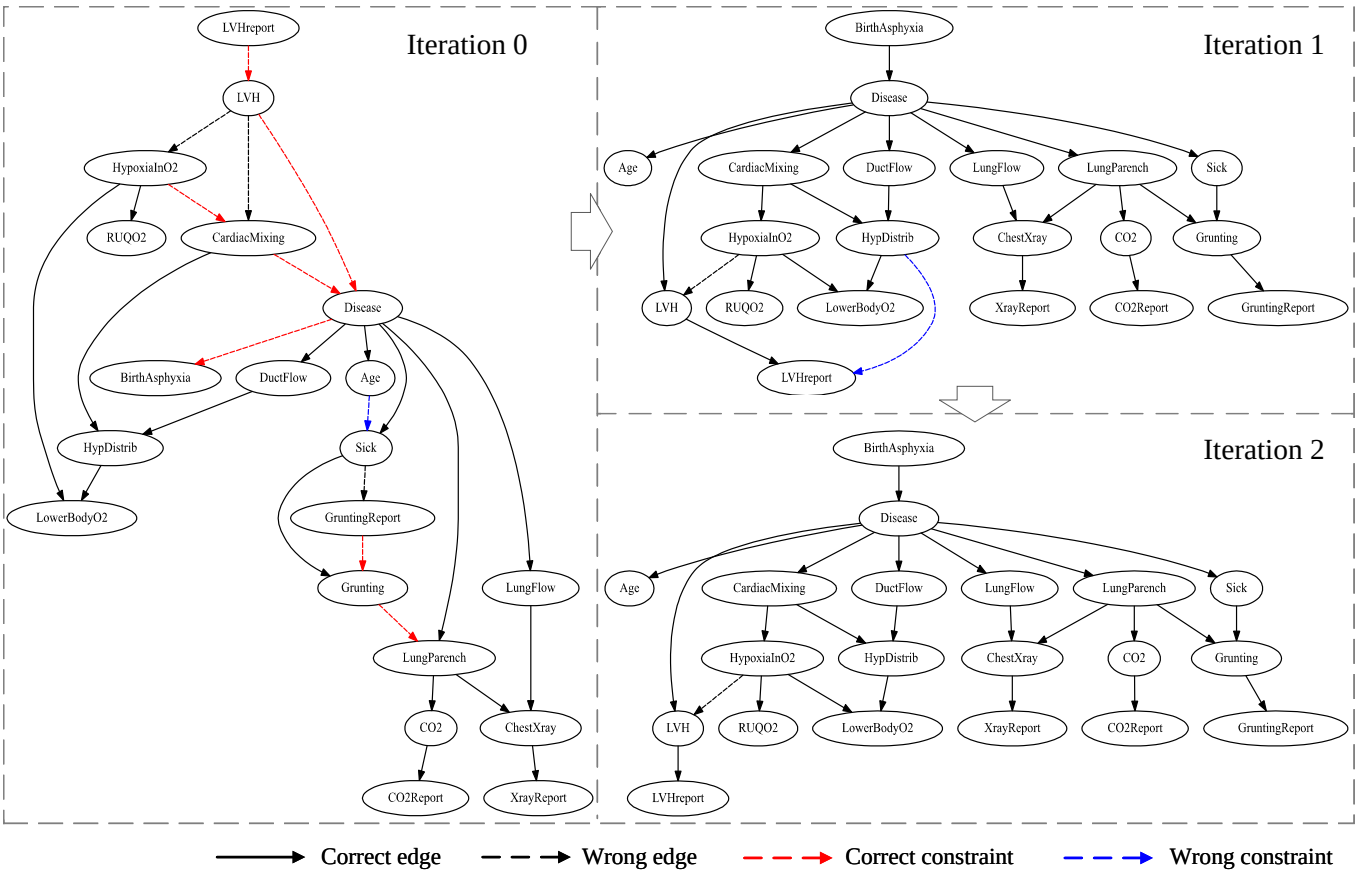


Fig. 3: Visualized process of HC-BDeu+ILS-CSL-hard on a set of observed data of *Child*, 2000 samples. The SHD of iterations are: 12 for Iteration 0, 3 for Iterations 1 and 2.

knowledge on them. It underscores the capability of discovering structures unrelated to prior constraints by integrating them. This phenomenon could be interpreted as the capability of aiding discovery of unknown causal mechanisms by the known causal knowledge.

VIII. Conclusions

This paper presents ILS-CSL, a framework that enhances causal discovery from data using Large Language Models (LLMs). ILS-CSL seamlessly incorporates LLM inference on the edges of the learned causal Directed Acyclic Graph (DAG), converting qualitative causal statements into precise edge-level prior constraints while effectively mitigating constraint errors stemming from imperfect prior knowledge. Comprehensive experiments across eight real-world datasets demonstrate the substantial and consistent improvement ILS-CSL brings to the quality of causal structure learning (CSL) outputs. Notably, ILS-CSL surpasses the existing separate way to guide CSL by applying LLM inferred causality as ancestral constraints, with a marked performance increase as the number of variables grows. This advancement underscores the promising application of the ILS-CSL framework in assistance of complex, real-world causal discovery tasks.

References

[1] J. Pearl, *Causality*. Cambridge university press, 2009.

[2] B. Ellis and W. H. Wong, "Learning causal bayesian network structures from experimental data," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 778–789, 2008.

[3] D. M. Chickering, "Learning bayesian networks is np-complete," *Learning from data: Artificial intelligence and statistics V*, pp. 121–130, 1996.

[4] N. K. Kitson, A. C. Constantinou, Z. Guo, Y. Liu, and K. Chobtham, "A survey of bayesian network structure learning," *Artificial Intelligence Review*, pp. 1–94, 2023.

[5] S. L. Morgan and C. Winship, *Counterfactuals and causal inference*. Cambridge University Press, 2015.

[6] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[7] E. Y.-J. Chen, Y. Shen, A. Choi, and A. Darwiche, "Learning bayesian networks with ancestral constraints," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[8] H. Amirkhani, M. Rahmati, P. J. Lucas, and A. Hommersom, "Exploiting experts' knowledge for structure learning of bayesian networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2154–2170, 2016.

[9] A. C. Constantinou, Z. Guo, and N. K. Kitson, "The impact of prior knowledge on causal structure learning," *Knowledge and Information Systems*, pp. 1–50, 2023.

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

[11] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.

[12] L. Chen, T. Ban, X. Wang, D. Lyu, and H. Chen, "Mitigating prior errors in causal structure learning: Towards llm driven prior knowledge," *arXiv preprint arXiv:2306.07032*, 2023.
[13] R. Tu, C. Ma, and C. Zhang, "Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis," *arXiv preprint arXiv:2301.13819*, 2023.
[14] S. Long, T. Schuster, A. Piché, S. Research *et al.*, "Can large language models build causal graphs?" *arXiv preprint arXiv:2303.05279*, 2023.
[15] 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," *arXiv preprint arXiv:2306.16902*, 2023.
[16] A. Vashishtha, A. G. Reddy, A. Kumar, S. Bachu, V. N. Balasubramanian, and A. Sharma, "Causal inference using llm-guided discovery," *arXiv preprint arXiv:2310.15117*, 2023.
[17] M. Willig, M. Zečević, D. S. Dhami, and K. Kersting, "Can foundation models talk causality?" *arXiv preprint arXiv:2206.10591*, 2022.
[18] H. Liu, R. Ning, Z. Teng, J. Liu, Q. Zhou, and Y. Zhang, "Evaluating the logical reasoning ability of chatgpt and gpt-4," *arXiv preprint arXiv:2304.03439*, 2023.
[19] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," *Advances in neural information processing systems*, vol. 21, 2008.
[20] J. Frohberg and F. Binder, "Crass: A novel data set and benchmark to test counterfactual reasoning of large language models," *arXiv preprint arXiv:2112.11941*, 2021.
[21] D. Heckerman, "A bayesian approach to learning causal networks," *arXiv preprint arXiv:1302.4958*, 2013.
[22] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social science computer review*, vol. 9, no. 1, pp. 62–72, 1991.
[23] E. V. Strobl, S. Visweswaran, and P. L. Spirtes, "Fast causal inference with non-random missingness by test-wise deletion," *International journal of data science and analytics*, vol. 6, pp. 47–62, 2018.
[24] D. Heckerman and D. Geiger, "Learning bayesian networks: a unification for discrete and gaussian domains," in *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, 1995, pp. 274–284.
[25] A. A. Neath and J. E. Cavanaugh, "The bayesian information criterion: background, derivation, and applications," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 2, pp. 199–203, 2012.
[26] C. Yuan, B. Malone, and X. Wu, "Learning optimal bayesian networks using a* search," in *Twenty-second international joint conference on artificial intelligence*, 2011.
[27] F. Trösser, S. de Givry, and G. Katsirelos, "Improved acyclicity reasoning for bayesian network structure learning with constraint programming," *arXiv preprint arXiv:2106.12269*, 2021.
[28] A. Li and P. Beek, "Bayesian network structure learning with side constraints," in *International Conference on Probabilistic Graphical Models*. PMLR, 2018, pp. 225–236.
[29] L. M. de Campos and J. G. Castellano, "Bayesian network learning algorithms using structural restrictions," *International Journal of Approximate Reasoning*, vol. 45, no. 2, pp. 233–254, 2007.
[30] G. F. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine learning*, vol. 9, pp. 309–347, 1992.
[31] R. T. O'Donnell, A. E. Nicholson, B. Han, K. B. Korb, M. J. Alam, and L. R. Hope, "Causal discovery with prior information," in *AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings 19*. Springer, 2006, pp. 1162–1167.
[32] 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 and Knowledge Discovery*, vol. 22, pp. 106–148, 2011.
[33] C. Lee and P. van Beek, "Metaheuristics for score-and-search bayesian network structure learning," in *Advances in Artificial Intelligence: 30th Canadian Conference on Artificial Intelligence, Canadian AI 2017, Edmonton, AB, Canada, May 16-19, 2017, Proceedings 30*. Springer, 2017, pp. 129–141.
[34] M. Scutari, C. E. Graafland, and J. M. Gutiérrez, "Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms," *International Journal of Approximate Reasoning*, vol. 115, pp. 235–253, 2019.

APPENDIX

A. Supplementary results on errors of LLM inference and prior constraints

In this section, we present the count of incorrect causal statements inferred by GPT-4 along with the erroneous prior constraints across various backbone algorithms and distinct observed data sizes of eight datasets. Figure 4 delineates the results pertinent to the hard constraining approaches, while Figure 5 elucidates those relevant to the soft constraining approaches.

B. Trends in Learned DAG Quality Over Iterations

This section outlines the iterative trends of scaled SHD (aiming for a decrease, denoted as SHD↓) and True Positive Rate (aiming for an increase, denoted as TPR↑) for various backbone algorithms across eight datasets, as depicted in Figure 6.

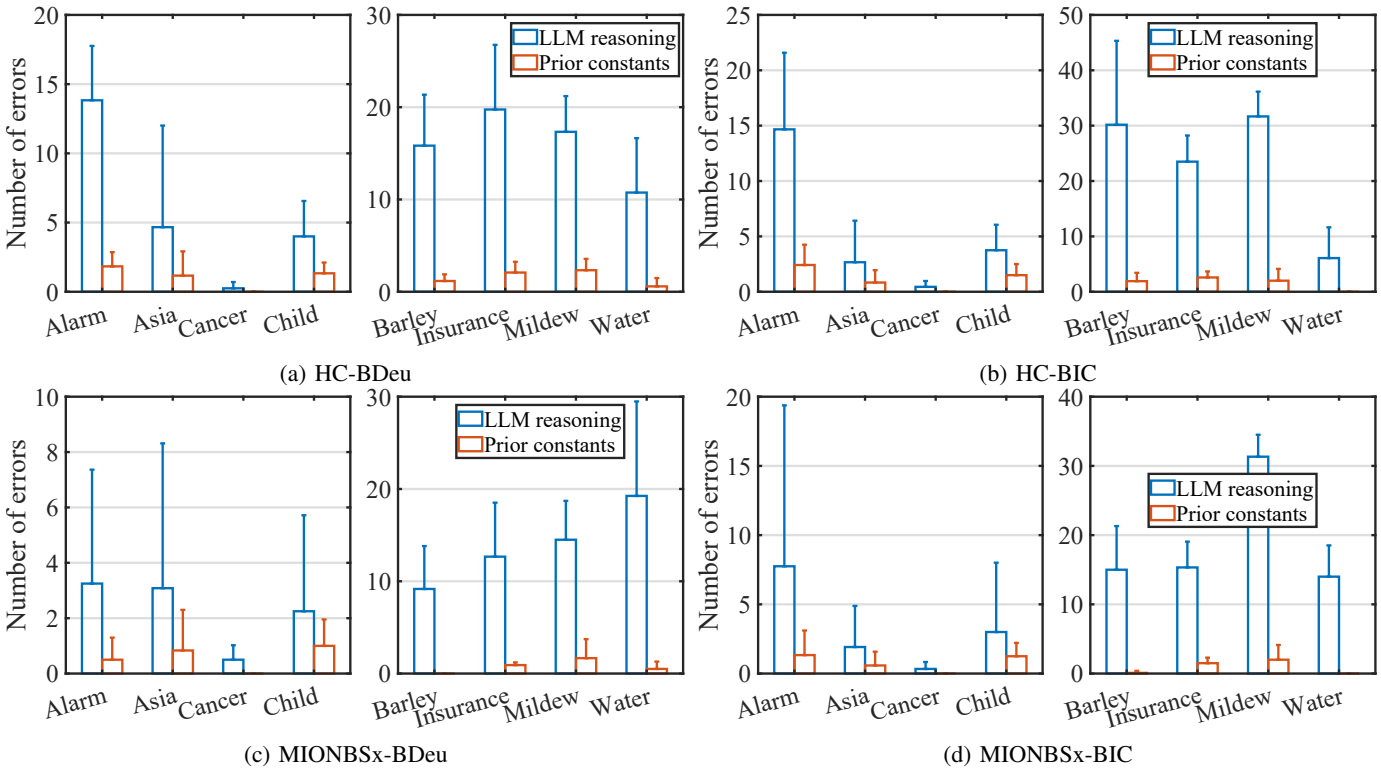
C. Trend of Constraints Derived from LLM Over Iterations

This section discusses the trend in the number of total and erroneous prior constraints derived from various backbone algorithms on eight datasets, as illustrated in Figure 7.

The following key observations emerge from the analysis:

- **Increasing Total Prior Constraints with Few Errors:** As the iterations progress, the number of total prior constraints sees a rise, while the increase in erroneous constraints is considerably smaller. This trend highlights the robust capability of ILS-CSL in generating high-quality, reliable constraints, enhancing the overall efficiency and reliability of the causal discovery process.
- **Occasional Decrease in Constraints:** Despite a general increase, some iterations exhibit a decrease in the number of prior constraints. This phenomenon is attributed to the same statistical artifact discussed in Appendix B. Some cases conclude in earlier iterations, leading to a varied set of statistical points across consecutive iterations, thereby affecting the total count of constraints.

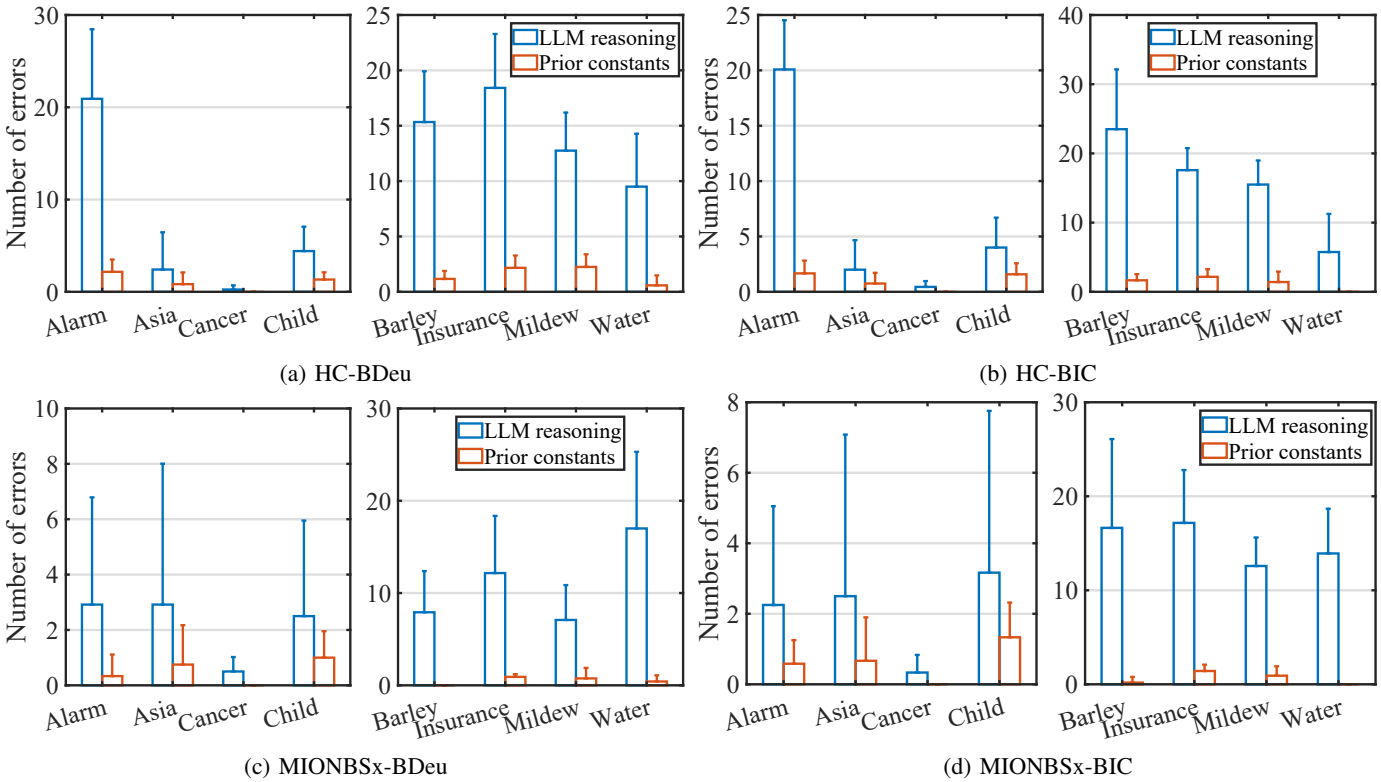
These observations further affirm the effectiveness of ILS-CSL in consistently generating high-quality constraints throughout the iterations.



(a) HC-BDeu (b) HC-BIC

(c) MIONBSx-BDeu (d) MIONBSx-BIC

Fig. 4: Number of erroneous LLM inference and prior constraints during ILS-CSL related to hard constraining approaches on various algorithms and datasets.



(a) HC-BDeu (b) HC-BIC

(c) MIONBSx-BDeu (d) MIONBSx-BIC

Fig. 5: Number of erroneous LLM inference and prior constraints during ILS-CSL related to soft constraining approaches on various algorithms and datasets.

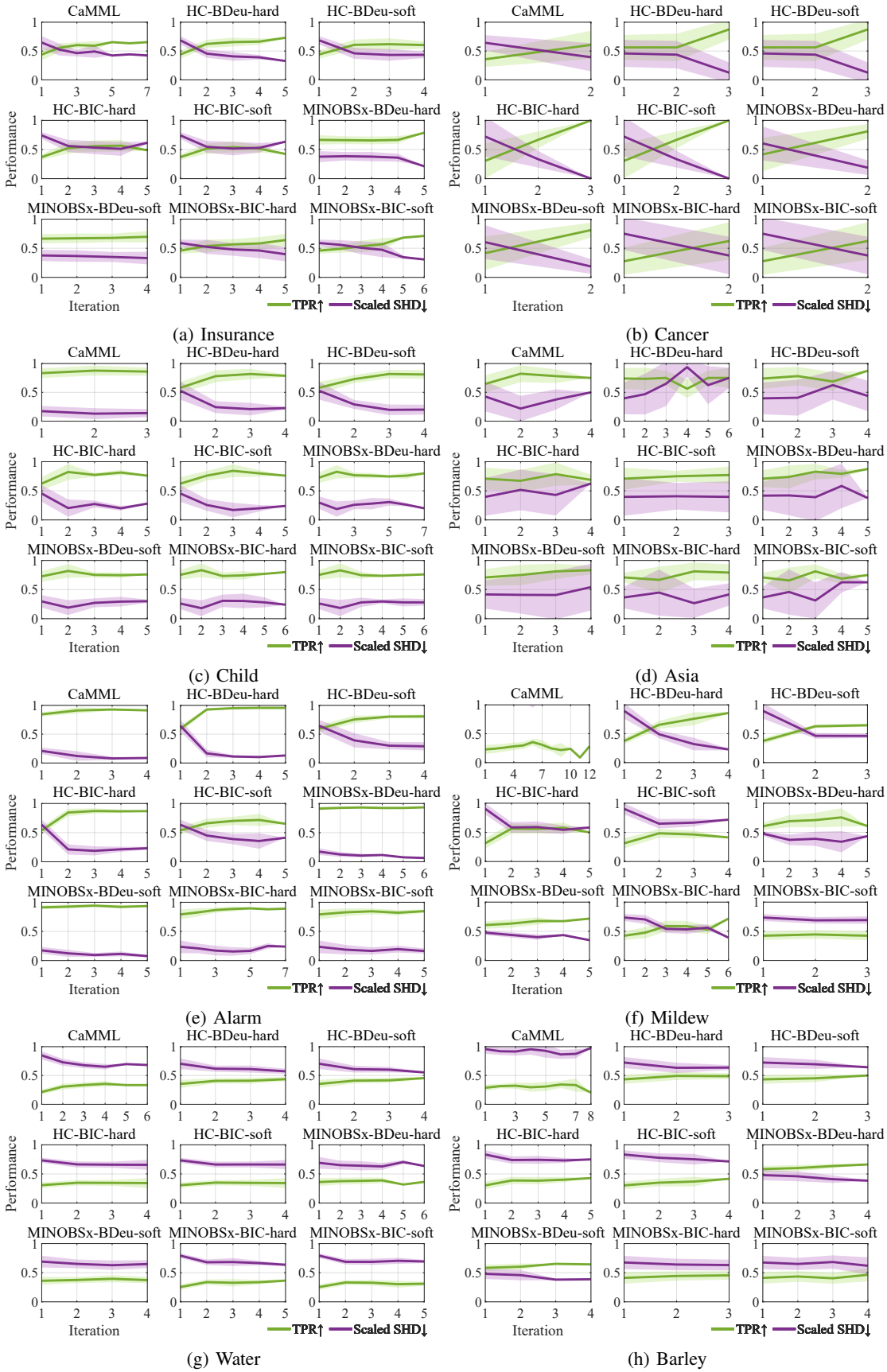


Fig. 6: TPR↑ (green line) and scaled SHD↓ (purple line) alongwith derivations (colored area) in ILS-CSL with various algorithms on various datasets.

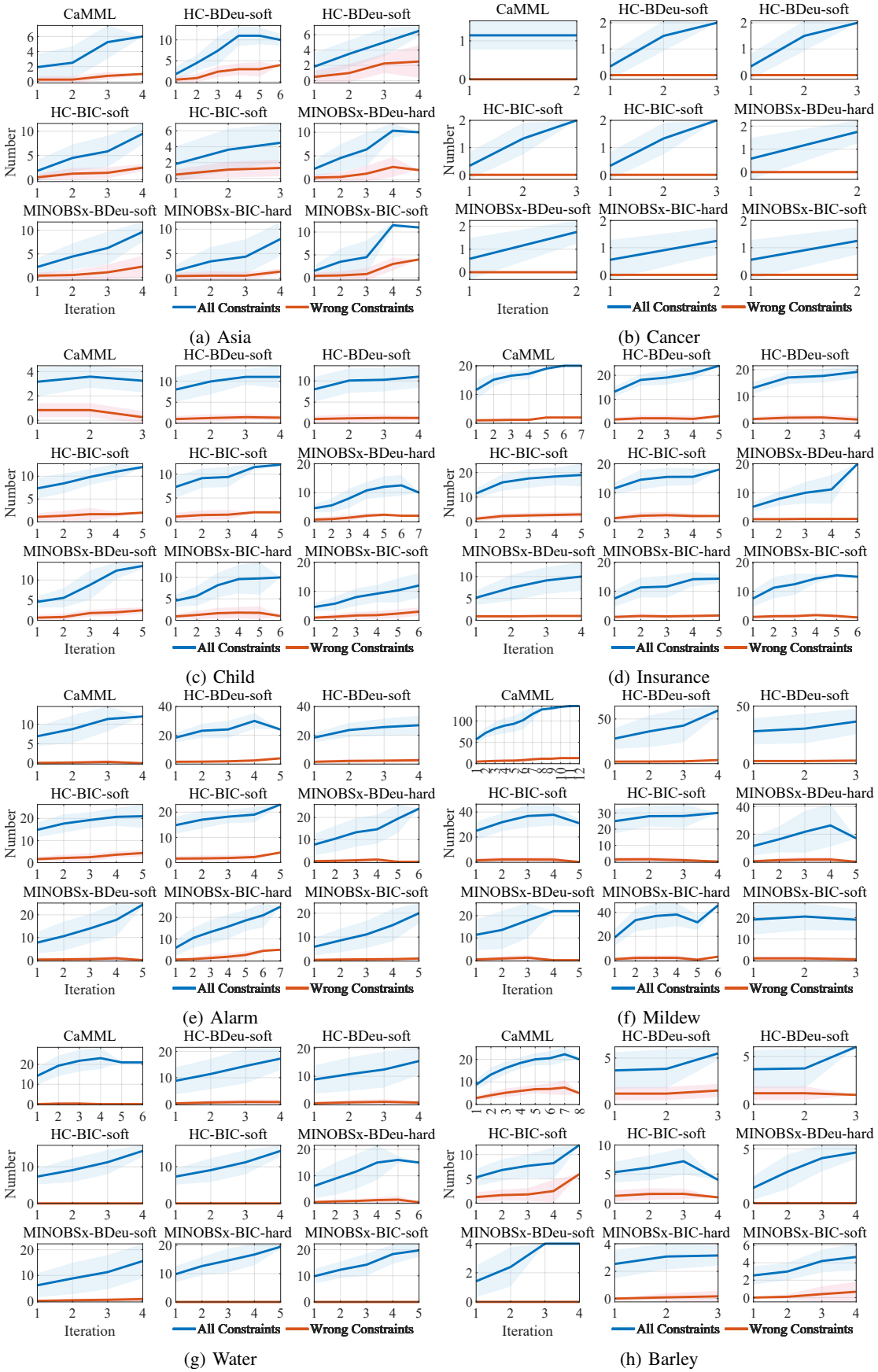


Fig. 7: Number of total (blue line above) and erroneous (red line below) prior constraints along with derivations (colored area) in ILS-CSL with various algorithms on various datasets.