CAUSAL STRUCTURE LEARNING SUPERVISED BY LARGE LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery from observational data is pivotal for deciphering complex relationships. While Causal Structure Learning (CSL) aims to extract causal Directed Acyclic Graphs (DAGs), its efficacy is hampered by the expansive DAG space and data sparsity. The advent of Large Language Models (LLMs) presents a novel avenue, given their aptitude in causal reasoning, thereby constraining CSL with knowledge-based causal inference. A pioneering study integrated LLMs into CSL, achieving notable results in several real-world DAGs. Yet, it faced pitfalls such as erroneous LLM inferences and the inefficacy of ancestral constraints. In response, we introduce the Iterative LLM Supervised CSL (ILS-CSL) framework. This approach seamlessly merges LLM-based causal inference with CSL, iteratively refining the causal DAG based on LLM feedback. Given LLM's shortness in distinguishing indirect causality form the direct, ILS-CSL is still capable to offer constraints on direct causality that are more powerful than the indirect, by integrating statistical dependencies indicated by data. Moreover, the prior errors are significantly reduced while using identical LLM resources. Our evaluations on eight real-world datasets confirm ILS-CSL's dominance, establishing a new benchmark in CSL performance.

1 INTRODUCTION

Causal discovery from observational data is pivotal in understanding intricate relationships across various domains. A primary method, Causal Structure Learning (CSL), seeks to derive a causal Directed Acyclic Graph (DAG)[1] from observed data (Pearl, 2009). However, the super-exponential growth of the DAG space introduces formidable challenges (Chickering, 1996). Exact algorithms struggle with scalability, while scalable approximate algorithms are prone to local optima, yielding low-quality structures (Kitson et al., 2023). Additionally, the typical sparse real-world data is insufficient for accurately discerning causal DAGs (Morgan & Winship, 2015). Moreover, the appropriate causal structures cannot be fully determined solely through statistical analysis of data, leading to confusion about the direction of causality (Chickering, 2002).

Given the inherent limitations of purely data-driven CSL, the integration of prior knowledge to constrain specific structures has been explored (Chen et al., 2016; Amirkhani et al., 2016). While promising, this approach has been limited by the high costs and time associated with expert input (Constantinou et al., 2023). 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 (Kıcıman et al., 2023; Nori et al., 2023; Chen et al., 2023).

A most recent work pioneers the integration of LLMs into CSL (Ban et al., 2023). Given the limitation of LLMs in specifying edge-level structures (Tu et al., 2023; Long et al., 2023) due to the shortness in distinguishing direct causality from the the indirect, the study utilizes the LLM inferred causal statements to constrain the *existence of paths* in CSL. The authors reach significant state-of-the-art performance transcending the pure data-based CSL in four out of eight real-world causal DAGs. However, their method falls short in the rest datasets bothered by the following issues:

[1]The causal DAG is the DAG in which each link represents direct functional relationship among the corresponding variables.

1. **Erroneous Inference:** LLMs often infers extraneous causal relationships, introducing erroneous structural constraints harming the CSL, please refer to Table 4 in Appendix A.1.
2. **Ancestral Constraints:** The used ancestral constraints, derived from LLM outputs, are less effective[2] and more risk-prone[3] than edge specifications (Li & Beek, 2018), especially as the number of variables increases, please refer to Table 5 in Appendix A.1.
3. **Batching Complexity:** They simultaneously query the LLM for multiple causal relationships to enhance efficiency, which, however, compromises inference quality and nuance.

In response to the challenges, we introduce an iterative LLM supervised CSL framework (ILS-CSL). Unlike the previous method that used LLMs to guide CSL in a separate manner, our method seamlessly integrates LLM-based causal inference and CSL. Concretely, we employ the LLM to verify the correctness of edges in the learned causal DAG and adapt the subsequent round of CSL to rectify mistakes identified in the prior round. The iterative process terminates once there's no further discord between LLM-based inferences and data-based CSL within the established causal skeleton. The main advantage of our method is summarized as follows:

1. **Powerful Structural Constraints:** By integrating direct dependencies (skeleton) indicated by data, ILS-CSL transforms the causal inferences made by LLMs into precise structural constraints, explicitly indicating the *edge existence or absence*. The edge-level constraint is more powerful than its path-level counterpart in improving CSL.
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 4.2 for detailed estimation.
3. **Efficient Causal Inference by LLM:** The number of pairwise variables for inference is reduced from $\binom{N}{2}$ to $O(N)$ with N denoting the node count[4]. It allows for individual conversations dedicated to inferring nuanced causality between pairwise variables.

We experiment on eight real-world datasets used in the initial LLM-driven CSL study (Ban et al., 2023), demonstrating consistent enhancement of data-based CSL algorithms by our framework across diverse scoring functions and search strategies. Moreover, ILS-CSL notably outperforms the first work, particularly with increased variable counts. The results highlight its substantial potential for aiding complex, real-world causal discovery tasks.

2 PRELIMINARIES

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

Causal structure learning. CSL seeks to uncover a directed acyclic graph (DAG) from observational data over a set of random variables. This DAG aims to fully characterize the conditional dependencies between variables and to be minimal (Pearl, 2009). 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 as follows:

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

$$\max_{\mathcal{G}} \sigma(\mathcal{G}; \mathbf{D}) = \sum_{i=1}^{n} \mathcal{L}_\sigma(X_i \mid X_{\mathrm{pa}(i)}; \mathbf{D}),\ X_{\mathrm{pa}(i)} = \{X_j \mid X_j \to X_i \in E(\mathcal{G})\} \text{ s.t. } \mathcal{G} \in \mathrm{DAG} \tag{2}$$

[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 $O(N)$.

Equations (1) and (2) define the CSL task in the context of its two main solutions: constraint-based and score-based methods. 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 DAG constraints, albeit not all (Spirtes & Glymour, 1991; Strobl et al., 2018). 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, $\mathcal{L}_\sigma(X_i \mid X_{\mathrm{pa}(i)}; \mathbf{D})$, where $X_{\mathrm{pa}(i)}$ denotes the set of parent nodes of X_i (Heckerman & Geiger, 1995; Neath & Cavanaugh, 2012). 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 (Yuan et al., 2011; Trösser et al., 2021). The best-scored DAG of the searched orderings is then selected as the output.

Due to the ability of score-based methods to orient all edges, thereby facilitating the determination of edge correctness, they serve as the foundational CSL algorithm in this paper.

Prior Constraints on Structures. Structural constraints play a pivotal role in improving the discovery of causal structures. The most prevalent among these constraints include (Li & Beek, 2018):

- **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 (Kıcıman et al., 2023; Ban et al., 2023; Tu et al., 2023), LLMs lack the ability to accurately specify direct causality, often confusing it with indirect causality or non-causal correlations. Please refer to Appendix A.5.2 for empirical estimation.

Algorithm 1: LLM supervised CSL

Require: Observed data, $\mathbf{D}$; Textual descriptions, $\mathbf{T}$
Ensure: Causal DAG, $\mathcal{G}$

```
Initialize the set of structural constraints, λ ← {}
repeat
    G ← arg max_G σ(G; D), s.t. G ∈ DAG, G ⊨ λ
    for X_i → X_j ∈ E(G) do
        c ← LLM infers causality between X_i and X_j based on T
        if c is X_i ← X_j then
            λ ← λ ∪ {X_j → X_i}
        end if
        if c is X_i ↮ X_j then
            λ ← λ ∪ {X_i ↛ X_j, X_j ↛ X_i}
        end if
    end for
until no new constraints are added
return 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 (de Campos & Castellano, 2007). Conversely, the soft approach strikes a balance between honoring prior constraints and the associated score costs (Amirkhani et al., 2016). This often involves adjusting the scoring function to $\sigma(\mathcal{G}; \mathbf{D}) + b(\mathcal{G}; \lambda)$, where a prior probability $P(\lambda)$ 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.

3 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.

1) 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 (Kıcıman et al., 2023), 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.

2) Constrain CSL: 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.

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.

Hard approach: We apply λ by pruning the space of local structures:

$$L(X_i; \lambda) \leftarrow \{X_{pa(i)} \mid K(i) \subseteq X_{pa(i)} \subseteq C(i)\} \tag{3}$$

$$K(i) = \{X_j \mid X_j \to X_i \in \lambda\},\ C(i) = X \setminus \{X_j \mid X_j \nrightarrow X_i \in \lambda\} \setminus \{X_i\} \tag{4}$$

where $\mathcal{L}_\sigma(\cdot)$ is the score of the local structure of X_i and its parent nodes $X_{pa(i)}$. The pruned space of local structures is taken as input for the search method $\mathcal{M}$:

$$\mathcal{M} : \max_{X_{pa(i)}} \sum_i^n \mathcal{L}_\sigma(X_i \mid X_{pa(i)}; \mathbf{D}),\ \text{s.t.}\, \mathcal{G} \in \text{DAG},\ X_{pa(i)} \in L(X_i; \lambda) \tag{5}$$

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

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

$$\sigma'(G; \mathcal{D}, \lambda) = \sum_i^n \mathcal{L}_\sigma(X_i \mid X_{pa(i)}; \mathbf{D}) + \mathcal{L}_b(X_i, X_{pa(i)}; \lambda) \tag{6}$$

$$\begin{aligned} \mathcal{L}_b(X_i, X_{pa(i)}; \lambda) = &\sum_{X_j \to X_i \in \lambda} \mathbb{I}_{X_j \in X_{pa(i)}} \log P(\lambda) + \mathbb{I}_{X_j \notin X_{pa(i)}} \log\left(1 - P(\lambda)\right) + \\ &\sum_{X_j \nrightarrow X_i \in \lambda} \mathbb{I}_{X_j \in X_{pa(i)}} \log\left(1 - P(\lambda)\right) + \mathbb{I}_{X_j \notin X_{pa(i)}} \log P(\lambda) \end{aligned} \tag{7}$$

This formulation is grounded in the decomposability of edge constraints. A detailed derivation can be found in Appendix A.2. $\mathbb{I}_{\texttt{condition}}$ is the indicator function, which takes the value 1 if the `condition` is true and 0 otherwise. $P(\lambda)$ is the prior confidence, a hyper-parameter. Then search method M optimizes the modified score:

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

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.

4 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.

4.1 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:

$$\max_{X_{pa(i)}: X_{pa(i)} \subseteq \{X \mid X \prec X_i \text{ in } O\}} \sigma(X_i \mid X_{pa(i)}), \ \forall i \in \{1, 2, ..., N\}$$

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[5]:

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.

4.2 Estimation of Prior Error Counts

Our objective is to estimate and juxtapose the number of erroneous constraints in our framework against those stemming from a full inference on all pairwise variables, an intuitive strategy that the first study (Ban et al., 2023) employs. Note that even if the authors reduce the number of valid causal statements by conducting $\binom{N}{2}$ pairwise causal inferences with a single prompt, the essential idea of their method is to fully constrain the existence of causality inferred between each pair of variables.

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

[5]Please refer to Appendix A.3 for proof.

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 $X_2 \rightarrow X_1$, it's an instance of extra causality.
4. Missing Direct Causality (p_m^d): If an edge $X_1 \rightarrow X_2$ or $X_2 \rightarrow X_1$ 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:

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

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 $X_1 \rightarrow X_2$ 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 A.4 for details. The results are present as:

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

And then we have:

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

For a more in-depth analysis of the good resistance to erroneous prior constraints of ILS-CSL, please refer to Appendix A.5.2.

Table 1: 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

5 EXPERIMENTS

We conduct experiments to address the following research questions:

RQ1: As a model-agnostic framework, can ILS-CSL enhance data-based CSL baselines and outperform the existing LLM-driven CSL method?
RQ2: Across varied scoring functions and search strategies, can ILS-CSL consistently improve the quality of discovered causl structures? How about the influence of strategies to apply constraints?
RQ3: Is ILS-CSL capable of minimizing erroneous prior constraints that arise from imperfect LLM causal inferences?
RQ4: How does the process, where LLM supervises causal discovery, unfold in detail?

Table 2: 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 (Ban et al., 2023), where the authors do not report derivations.

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 study by (Ban et al., 2023). Dataset specifics are provided in Table 1. For backbone CSL algorithms, we adopt the same MINOBSx (BDeu score) (Li & Beek, 2018) and CaMML (MML score) (O'Donnell et al., 2006) algorithms as the aforementioned study, 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 (Heckerman & Geiger, 1995), and search algorithms, including HC (Gámez et al., 2011) and MINOBSx (Lee & van Beek, 2017).

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 the study (Li & Beek, 2018) and used in the LLM-driven CSL by Ban et al. (2023). The repository offers datasets in two distinct sample sizes for each DAG, as detailed in Table 1. 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) (Scutari et al., 2019). This metric is defined as the SHD normalized by the total number of edges in the true causal DAG.

5.1 COMPARATIVE PERFORMANCE OF CSL BASED ON PURE DATA, SEPERATE LLM PRIOR AND ILS-CSL (RQ1)

We compare the performance of MINOBSx (BDeu) and CaMML that are used in the separate LLM prior-driven CSL approach proposed by (Ban et al., 2023), referred to as sepLLM, and our proposed framework, termed ILS-CSL. This comparison is conducted using all the introduced observed data

[6]https://www.bnlearn.com/bnrepository/
[7]https://chat.openai.com/
[8]https://github.com/andrewli77/MINOBS-anc/tree/master/data/csv

across eight datasets. The results, presented in terms of scaled SHD (where a lower value is preferable), are detailed in Table 2. The difference between scaled SHD of data-based (Δ_{data}) and LLM-driven (Δ_{LLM}) CSL is also reported, by calculating $(\Delta_{\text{LLM}} - \Delta_{\text{data}})/\Delta_{\text{data}}$. Please refer to Appendix A.5.1 for the ranking of the methods and more detailed discussions.

Result observation. 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 6 in Appendix A.4). 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 3: 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%

5.2 Effect of ILS-CSL across different scores and 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 3.

Result observation. 1) Nearly all scenarios showcase an enhancement, underscoring the impactful role of ILS-CSL in bolstering CSL performance across a diverse range of datasets and conditions.
2) ILS-CSL notably amplifies the performance of HC more than MINOBSx. This enhancement is so pronounced that HC, when integrated with ILS-CSL, surpasses MINOBSx in performance, despite HC's inherently lower baseline performance. This observation highlights the substantial potential of ILS-CSL to markedly boost the performance of search algorithms, even enabling those with initially

lower effectiveness to outperform more advanced algorithms.
3) While in some cases ILS-CSL elevates BIC-based CSL to outshine BDeu baselines, this is not a universal occurrence. This inconsistency underscores the influence of the employed scoring function on the effectiveness of ILS-CSL, emphasizing the critical role the scoring function plays within a prior knowledge-driven CSL framework. Despite this, BIC-based CSL with ILS-CSL is greatly improved in ranking, see Appendix A.5.1 for the ranking of these methods along with further analysis.
4) 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 Ban et al. (2023), where the soft approach fared better due to the lower quality of prior constraints derived from full inference. This comparison further highlights the tolerance to imperfect LLM inference brought by ILS-CSL.

5.3 Assessment of Erroneous 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 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.

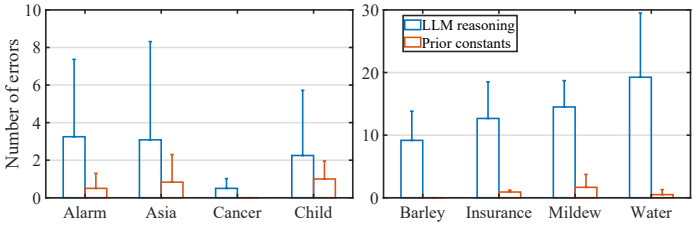


Figure 1: Erroneous LLM inference and erroneous specified edge constraints of MINOBSx-BDeu+ILS-CSL-hard.

Our observations highlight a substantial reduction in the errors of specified edge constraints compared to erroneous LLM inference. This is attributed to the strategy of only imposing constraints on causality that is inconsistent with what has been learned. This effect is grounded in the notable precision of LLMs in negating additional causality or rectifying the direction of reversed causality, which reduces the risk of their inaccuracy in pinpointing direct causality, please refer to Appendix A.5.2 for estimation of precision of different types of causality inferred by LLM.

It's important to note that various backbone CSL methods affect the output of causal DAGs, leading to a diverse range in the number of erroneous prior errors and LLM inferences. For a more comprehensive set of results, refer to Appendix A.5.3.

5.4 Iterative Details of ILS-CSL (RQ4)

We unfold details of ILS-CSL by 1) presenting the iterative trend of SHD with various backbone algorithms on the eight datasets, see Appendix A.6.3, 2) evaluating the trend of constraint numbers and errors in iteration, with results reported in Appendix A.6.3, and 3) visualize the evolution of causal DAGs in ILS-CSL for in-depth understanding of the process supervising causal discovery by LLM, please refer Appendix A.6.1.

6 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

Hossein Amirkhani, Mohammad Rahmati, Peter JF Lucas, and Arjen Hommersom. Exploiting experts' knowledge for structure learning of bayesian networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2154–2170, 2016.

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

Eunice Yuh-Jie Chen, Yujia Shen, Arthur Choi, and Adnan Darwiche. Learning bayesian networks with ancestral constraints. *Advances in Neural Information Processing Systems*, 29, 2016.

Lyuzhou Chen, Taiyu Ban, Xiangyu Wang, Derui Lyu, and Huanhuan Chen. Mitigating prior errors in causal structure learning: Towards llm driven prior knowledge. *arXiv preprint arXiv:2306.07032*, 2023.

David Maxwell Chickering. Learning bayesian networks is np-complete. *Learning from data: Artificial intelligence and statistics V*, pp. 121–130, 1996.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Anthony C Constantinou, Zhigao Guo, and Neville K Kitson. The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, pp. 1–50, 2023.

Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9:309–347, 1992.

Luis M de Campos and Javier G Castellano. Bayesian network learning algorithms using structural restrictions. *International Journal of Approximate Reasoning*, 45(2):233–254, 2007.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

Jörg Frohberg and Frank Binder. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models. *arXiv preprint arXiv:2112.11941*, 2021.

José A Gámez, Juan L Mateo, and José M Puerta. Learning bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. *Data Mining and Knowledge Discovery*, 22:106–148, 2011.

David Heckerman and Dan Geiger. Learning bayesian networks: a unification for discrete and gaussian domains. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pp. 274–284, 1995.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

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

Neville Kenneth Kitson, Anthony C Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. A survey of bayesian network structure learning. *Artificial Intelligence Review*, pp. 1–94, 2023.

Colin Lee and Peter 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*, pp. 129–141. Springer, 2017.

Andrew Li and Peter Beek. Bayesian network structure learning with side constraints. In *International Conference on Probabilistic Graphical Models*, pp. 225–236. PMLR, 2018.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*, 2023.

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

Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Rodney T O'Donnell, Ann E Nicholson, Bin Han, Kevin B Korb, M Jahangir Alam, and Lucas 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*, pp. 1162–1167. Springer, 2006.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, 115:235–253, 2019.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

Eric V Strobl, Shyam Visweswaran, and Peter L Spirtes. Fast causal inference with non-random missingness by test-wise deletion. *International journal of data science and analytics*, 6:47–62, 2018.

Fulya Trösser, Simon de Givry, and George Katsirelos. Improved acyclicity reasoning for bayesian network structure learning with constraint programming. *arXiv preprint arXiv:2106.12269*, 2021.

Ruibo Tu, Chao Ma, and Cheng Zhang. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. *arXiv preprint arXiv:2301.13819*, 2023.

Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality? *arXiv preprint arXiv:2206.10591*, 2022.

Changhe Yuan, Brandon Malone, and Xiaojian Wu. Learning optimal bayesian networks using a* search. In *Twenty-second international joint conference on artificial intelligence*, 2011.

A APPENDIX

A.1 RELATED WORK

LLM-based causal inference. The majority of current research on LLM-driven causal inference primarily focuses on evaluating the capabilities of LLMs themselves (Willig et al., 2022; Liu et al., 2023). For instance, this work (Long et al., 2023) assesses LLMs by probing their proficiency in generating simple causal structures for specific variable sets, typically comprising 3-4 variables. In a more domain-specific study, this study (Tu et al., 2023) tasks LLMs with discerning causal structures from variables pertinent to medical pain diagnosis. However, the results from this endeavor were less than satisfactory.

Building on these insights, Kıcıman et al. (2023) channel their efforts into devising more effective prompts, aiming to enhance LLM performance in discerning causal structures, particularly within the medical pain diagnosis dataset. However, the quality of the causal DAG that LLM outputs still has gap with that by data-based algorithms. Despite this, their research broadens the scope to evaluate LLMs across a diverse set of causal tasks. Notably, they report significant accuracy in tasks like pairwise causal discovery (Hoyer et al., 2008) and counterfactual reasoning (Frohberg & Binder, 2021). It's worth noting that, unlike causal structure learning, these tasks do not differentiate between direct and indirect relationships.

Integration of LLM and data-based CSL. A recent work first introduces LLM into the framework of data-driven CSL (Ban et al., 2023). Built upon the fact that LLMs are short in distinguishing indirect causality from the direct one, the authors apply ancestral constraints according to the statements of existence of causality between pairwise variables made by LLM. However, they acknowledge that LLM can conflate correlation with causality, leading to erroneous constraints on extra paths.

To reduce the harm of extra ancestral constraints and to enhance the efficiency of LLM inference, the authors provide the LLM with all investigated variables and prompt for the most confident causal statements, to reduce the number of causal statements and improves accuracy as possible. Yet, this approach significantly complicates the causality inference process. The LLM faces immense challenges in delivering as precise and comprehensive causal analyses as individual prompt for each pairwise variables. As the number of variables increases, the accuracy of the inferred causal statements diminishes, as evidenced in Table 4.

Furthermore, while ancestral constraints prove effective for small-scale DAGs, their efficacy wanes with an increasing number of variables. This decline can be attributed to the heightened likelihood of erroneous edges being incorporated into the recovered path, especially as inter-variable correlations become more intricate. This observation is further corroborated by the structural hamming distance (SHD) metrics presented by the authors in Table 5. Instances where the integration of GPT-4-driven ancestral constraints adversely impacted CSL are highlighted in gray. Notably, the performance of ancestral constraint-based CSL deteriorates in datasets like *Child*, *Insurance*, and *Alarm*, even when the GPT-4-driven prior constraints are deemed high-quality.

Table 4: Correct and total count of GPT-4 inferred causal statements reported by Ban et al. (2023).

Dataset	Cancer	Asia	Child	Insurance	Alarm	Water	Mildew	Barley
Correct / Total	5 / 5	9 / 9	8 / 10	10 / 10	20 / 21	9 / 15	5 / 8	17 / 24

Table 5: SHD↓ by data-based CSL and GPT-4 driven prior constraints reported by Ban et al. (2023).

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

*The bold SHD is the best performance in each dataset.

A.2 Derivation of Equations (6) and (7)

In this section, we derive a scoring function for the DAG $\mathcal{G}(X, E(\mathcal{G}))$ using data $\mathbf{D}$ and prior constraints, denoted as $\lambda : <\mathbf{R}, \mathbf{\Pi}>$. The set $\mathbf{R} = \{r_1, r_2, \cdots, r_m\}$ comprises edge variables on m pairwise variables, where $r_i \in \{\rightarrow, \nrightarrow\}$. $\mathbf{\Pi} = \Pi_{i=1}^{m} P(r_i)$ 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 (Cooper & Herskovits, 1992) and BDe, BDeu (Heckerman & Geiger, 1995), 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).

A.3 Proof of Lemma 1

To substantiate Lemma 1, we begin by highlighting a characteristic inherent to wildly used scoring functions designed specifically for CSL: the regularization mechanism[9]. This mechanism penalizes the addition of edges in the causal DAG to reduce the model complexity. Without this regularization, the DAG could gravitate towards becoming a complete graph, which would be nonsensical. It synergizes with the evaluation based on likelihood probability inferred from observed data under a specific distribution to constitute a scoring function.

When assessing a local structure, adding new variables to the parent set will lead to a penalty on the local score $\mathcal{L}_\sigma(X_i \mid X_{pa(i)}; \mathbf{D})$ due to the regularization mechanism. This implies that indiscriminately augmenting variables to the parent set doesn't consistently enhance the local score.

Building on this understanding, let's consider a scenario where the optimal and the second-best solutions for $X_{pa(i)}$, represented as X_{opt} and X'_{opt}, comprise distinct variables. Specifically, there exist variables X_j and X_k such that X_j is in X'_{opt} but not in X_{opt}, and vice versa for X_k. If we exclude X_k from the potential parent set, the optimal solution shifts to X'_{opt}. Given that X_j is now in X'_{opt} but absent in X_{opt}, it's evident that the revised optimal solution isn't a subset of its predecessor. This observation solidifies the proof for Lemma 1.

Consider a basic scenario where X_i has two candidate parent variables, A and B. If the score of either A or B individually serving as the parent of X_i surpasses the score when A and B are both parents (due to complexity penalties), then the following inequality holds: $\mathcal{L}_\sigma(X_i \mid \{A\}), \mathcal{L}_\sigma(X_i \mid \{B\}) > \mathcal{L}_\sigma(X_i \mid \{A, B\}) > \mathcal{L}_\sigma(X_i \mid \emptyset)$. In this scenario, $\{A\}$ and $\{B\}$ emerge as the optimal and suboptimal solutions, corroborating the condition described earlier.

Finally, we illustrate a concrete example where forbidding specific edges aids in uncovering a missing true edge. We consider the BDeu scores for the first of six observed data sets on the *Asia* dataset, comprising 1000 samples[10]. The true parent set for the variable X_3 `(lung cancer)` is $\{X_2$ `(smoking)`$\}$.

The local scores are ranked as follows:

$$\begin{aligned} & \mathcal{L}_\sigma(X_3 \mid X_{pa(3),1}) > \cdots > \mathcal{L}_\sigma(X_3 \mid X_{pa(3),k}) > \mathcal{L}_\sigma(X_3 \mid \{X_2\}) > \cdots > \mathcal{L}_\sigma(X_3 \mid \emptyset) > \cdots \\ & X_{pa(3),j} \cap \{X_5\texttt{(either)}, X_6\texttt{(positive xray)}, X_7\texttt{(dyspnoea)}\} \neq \emptyset, j = 1, ..., k \end{aligned} \tag{19}$$

where $X_{pa(3),i}$ is the parent set that ranks i. Note that the ground truth ranks $k+1$, and the empty set ranks lower. The optimal parent set (ranked first) for X_3 is:
$X_{pa(3),1} = \{X_1$ `(tuberculosis)`$, X_5$ `(either)`$\}$.

[9]This property is also known as global consistency, as referenced in (Chickering, 2002).

[10]Available in supplementary codes.

Furthermore, each of the top k parent sets for X_3 contains at least one variable from the set: $\{X_5\texttt{(either)}, X_6\texttt{(positive xray)}, X_7\texttt{(dyspnoea)}\}$.

The optimal parent set X_{opt} does not include the ground truth X_2. Supervised by GPT-4, none of X_5, X_6, X_7 directly causes X_2 based on intuitive knowledge, thereby constraining the parent set of X_3 to exclude X_5, X_6, and X_7. The new optimal X'_{opt} under the constraints is the ground truth $\{X_2\}$ according to Equation (19). In this case, the missing true edge $X_2 \to X_3$ is recovered by forbidding three edges when supervising edges of the learned DAG.

A.4 Parameter Estimation in Section 4.2

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, $\gamma_2, z_1, z_2, z_3, P_{R|E}$.

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 3. 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 6.

Table 6: 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

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.

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 7.

Table 7: 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

A.5 Supplementary Experiments

This section provides additional results and a more detailed analysis to enhance and expand upon the conclusions related to research questions 1, 2 and 3 presented in the main text.

A.5.1 Ranking of the investigated CSL methods (RQ1 and RQ2)

To rank the investigated CSL methods, we utilize Friedman test (Friedman, 1937), a popular method used in the context of comparing the performances of different algorithms over multiple datasets.

Concretely, the algorithms are ranked for each dataset. If there are n algorithms, assign ranks from 1 to n. If two algorithms perform equally well, assign them the average of the ranks they would have received. Then, calculate the Friedman statistic (χ_F^2) using the formula:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left(\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right)$$

where N is the number of datasets, k is the number of algorithms, and R_j is the rank sum of the j-th algorithm across all datasets. On this basis, compare the calculated χ_F^2 value with the critical value from the chi-square distribution with $k-1$ degrees of freedom to determine the significance, or calculate the p-value.

We report the friedman ranking and p-values of methods of Tables 2 and 3, in Tables 8 and 9, respectively. The best and second ranking are highlighted with bold and underlined texts, repsectively.

Table 8: Friedman ranking of methods in Table 2, where p-value equals 2.7e-4.

Data-based CSL		SepLLM		ILS-CSL	
MINOBSx	CaMML	MINOBSx	CaMML	MINOBSx	CaMML
3.6	4.8	4.0	3.9	**1.9**	<u>2.9</u>

Table 9: Friedman ranking of methods in Table 3, where p-value equals 5.7e-19.

Data-based CSL				ILS-CSL							
				Soft				Hard			
BDeu		BIC		BDeu		BIC		BDeu		BIC	
MINOBSx	HC	MINOBSx	HC	MINOBSx	HC	MINOBSx	HC	MINOBSx	HC	MINOBSx	HC
7.0	10.1	9.8	11.2	<u>2.8</u>	5.3	6.2	8.0	**2.7**	3.4	4.9	6.6

Result observation. We observe a notable trend in the rankings of backbone CSL algorithms with ILS-CSL, as they consistently secure positions within the top two as shown in Table 8. Intriguingly, within the sepLLM framework, CaMML (employing soft constraints) surpasses MINOBSx (employing hard constraints). This trend is reversed within the ILS-CSL framework. This phenomenon can be attributed to two primary factors:

1. The sepLLM framework tends to generate more erroneous constraints compared to ILS-CSL. Consequently, the error-tolerant nature of soft constraints proves more beneficial within the sepLLM framework. In contrast, within the ILS-CSL framework, where constraints are of higher quality, hard constraints demonstrate greater significance.
2. The use of ancestral constraints within the sepLLM framework can inadvertently introduce erroneous edges along a path, a problem not encountered with the edge constraints utilized within the ILS-CSL framework.

In essence, even correct ancestral constraints can inadvertently introduce risk into the CSL process, leading to scenarios where soft constraints may outperform hard constraints. This highlights the critical importance of constraint quality and the potential pitfalls of ancestral constraints. ILS-CSL adeptly addresses these two pivotal aspects, ensuring enhanced and stable performance while utilizing the same LLM resources. This strategic approach fortifies the robustness and reliability of ILS-CSL in diverse CSL scenarios, reinforcing its utility as a valuable tool for accurate causal structure learning.

Table 9 predominantly reveals the following observations:

- Almost all the methods employing ILS-CSL rank higher than their data-based CSL counterparts, with the exception of HC-BIC+ILS-CSL-soft (ranking 8.0) and MIONBSx+BDeu (ranking 7.0). This underscores the superior performance enhancement brought about by ILS-CSL, transcending the specific characteristics of the underlying backbone algorithms.
- The scoring function is more important than the search strategy. It is observed that ILS-CSL, when used with different search methods but the same score, tends to have closer rankings compared to the usage with the same search method but different scores. This suggests that lower-performing search strategies are more substantially improved by ILS-CSL than their scoring counterparts. This phenomenon can be attributed to the fact that prior constraints can streamline the search process by narrowing down the search space, albeit having a limited impact on amending inaccurate statistical data analysis.
- The hard constraint method consistently outperforms the soft constraint approach. This finding aligns with previous conclusions, reinforcing the notion that ILS-CSL contributes high-quality edge constraints to the causal discovery process. This advantage underscores the significance of satisfying every constraint, which in the context of ILS-CSL, outweighs the potential risks associated with errors.

These insights collectively highlight the robust and versatile performance enhancement capabilities of ILS-CSL across various scenarios and configurations, affirming its value as a significant asset in the realm of causal structure learning.

A.5.2 Understanding ILS-CSL's Resistance to Erroneous Constraints (RQ3)

This section elucidates the ability of ILS-CSL to minimize prior errors by limiting LLM supervision to edge-level pairwise variables. We present the ratio of various real structures corresponding to all pairwise variables inferred by GPT-4. Table 10 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 directions) 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 different types of causal relationships in the ground truth:
Direct Edges: The ratio of cases where X_1 directly causes X_2 ($X_1 \to X_2$).
Reversed Edges: The ratio of cases where X_2 directly causes X_1 ($X_2 \to X_1$).
Indirect Paths: The ratio of cases where X_1 indirectly leads to X_2 ($X_1 \rightsquigarrow X_2$) without a direct edge ($X_1 \nrightarrow X_2$).
Reversed Indirect Paths: The ratio of cases where X_2 indirectly leads to X_1 ($X_2 \rightsquigarrow X_1$) without a direct edge ($X_2 \nrightarrow X_1$).

Table 10: 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

Not Reachable: The ratio of cases where X_1 and X_2 are not reachable from each other ($X_1 \not\rightsquigarrow X_2, X_2 \not\rightsquigarrow X_1$).

Observation and Analysis. The edge-level pairwise variables precision (light red cells) is notably high, significantly exceeding the precision on arbitrary variables for both qualitative and structural aspects. Analyzing potential errors of 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 7) and the strategy of specifying a prior constraint only when inconsistent.

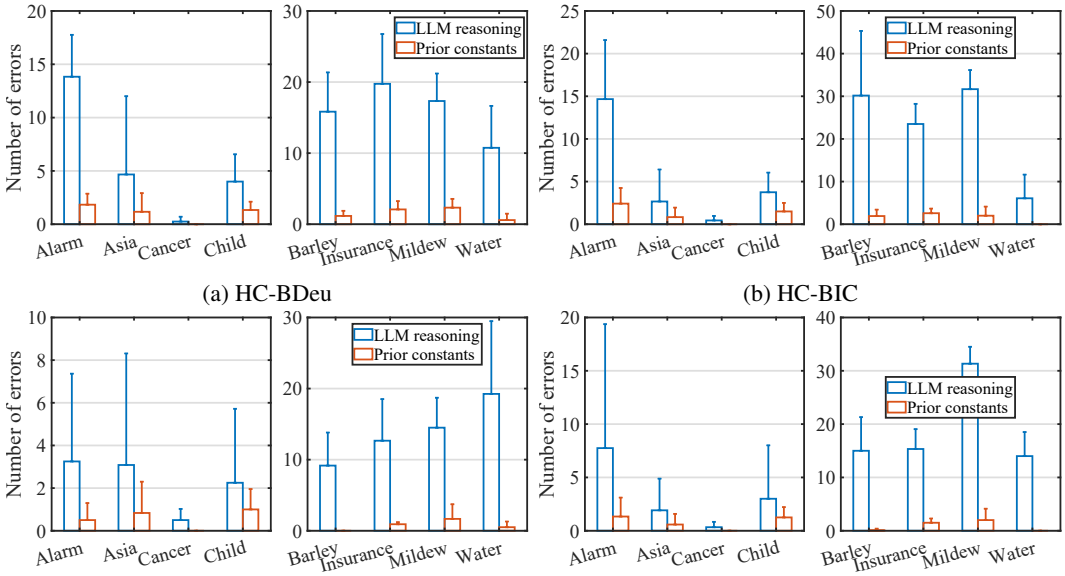
In essence, the primary limitation of LLM in causal inference is the confusion between direct causal relationships, indirect causality, and correlations, evidenced by the low overall qualitative and structural precision. This limitation hampers the performance of using LLM-derived existence on causality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

A.5.3 SUPPLEMENTARY RESULTS ON ERRORS OF LLM INFERENCE AND PRIOR CONSTRAINTS (RQ3)

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 2 delineates the results pertinent to the hard constraining approaches, while Figure 3 elucidates those relevant to the soft constraining approaches.

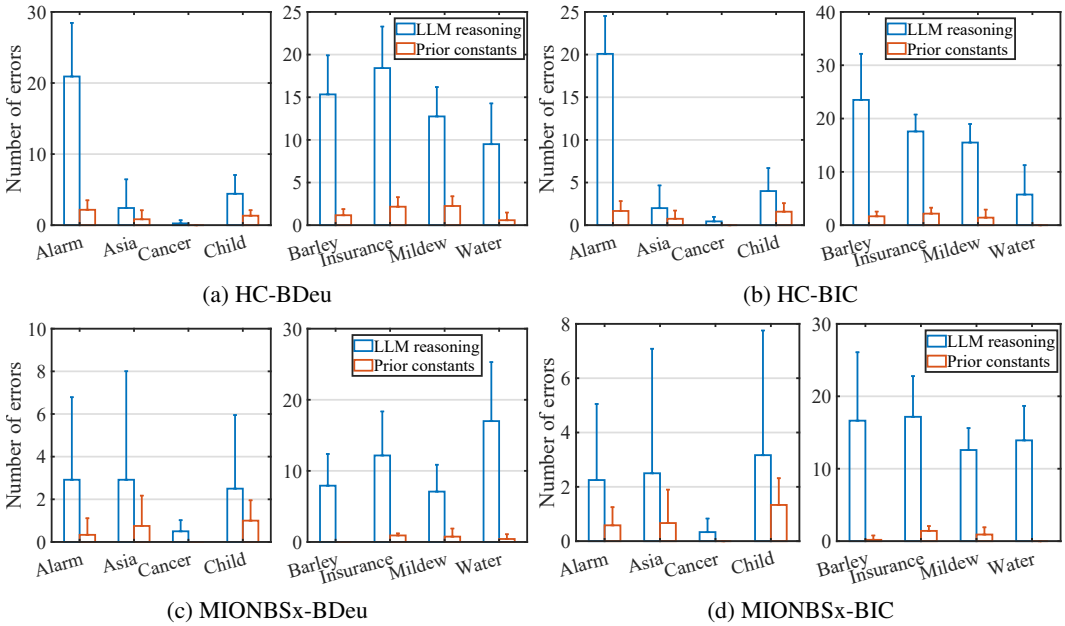
We observe that the number of prior constraints is much fewer than that of LLM inferences. Please refer Appendix A.5.2 for related analysis.



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

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

Figure 2: 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

Figure 3: Number of erroneous LLM inference and prior constraints during ILS-CSL related to soft constraining approaches on various algorithms and datasets.

A.6 Details and Visualization

In this section, we unfold the details of the iterative process of ILS-CSL for in-depth analysis, and visualize a representative example for intuitive understanding.

A.6.1 Visualization

We provide a visualized example to illustrate the effect of ILS-CSL in enhancing the quality of learned causal DAGs, including the improvement on the prior-independent structures (interpreted as the unknown causal mechanisms). The result of HC (BDeu) algorithm on *Child* dataset, 2000 samples, with hard constraining approach in ILS-CSL, is reported in Figure 4.

Key observations 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 direct refined edges by these priors are 5 $(7-2)$, while the reduced SHD is 8, meaning that 3 edges that are distinct from those in constraints are corrected without any prior 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 constraining the known knowledge on causality.

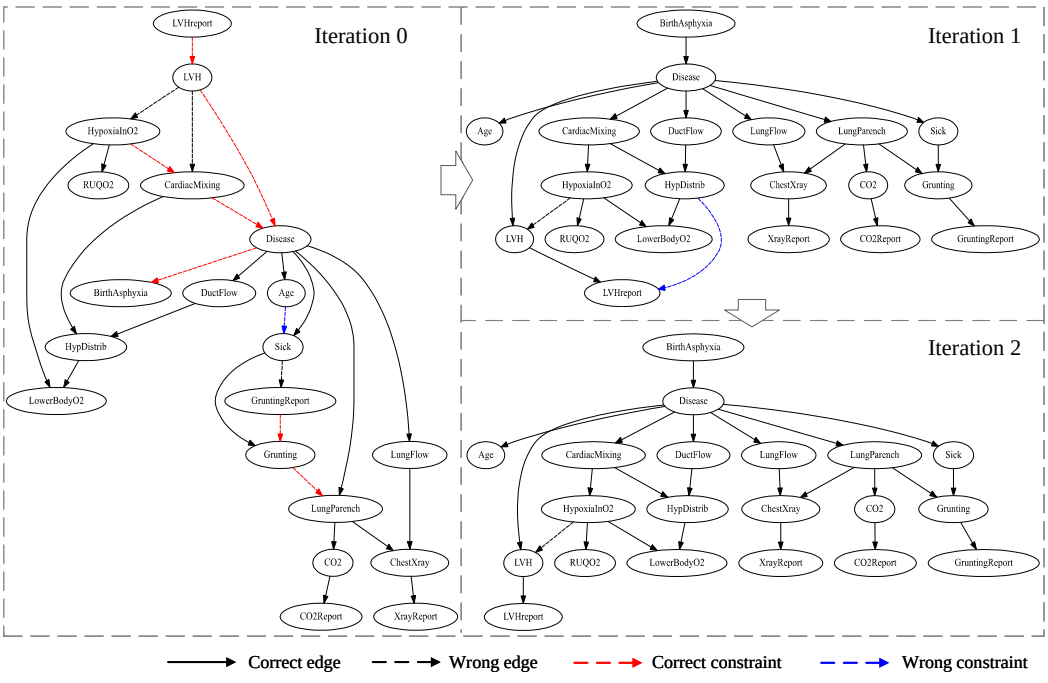


Figure 4: Visualized process of HC-BDeu+ILS-CSL-hard on a set of observed data of *Child*, 2000 samples. The SHD of iterations are: 12 for Iteration 0, 3 for Iterations 1 and 2.

A.6.2 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 5. 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*.

Observations: 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 5, 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.

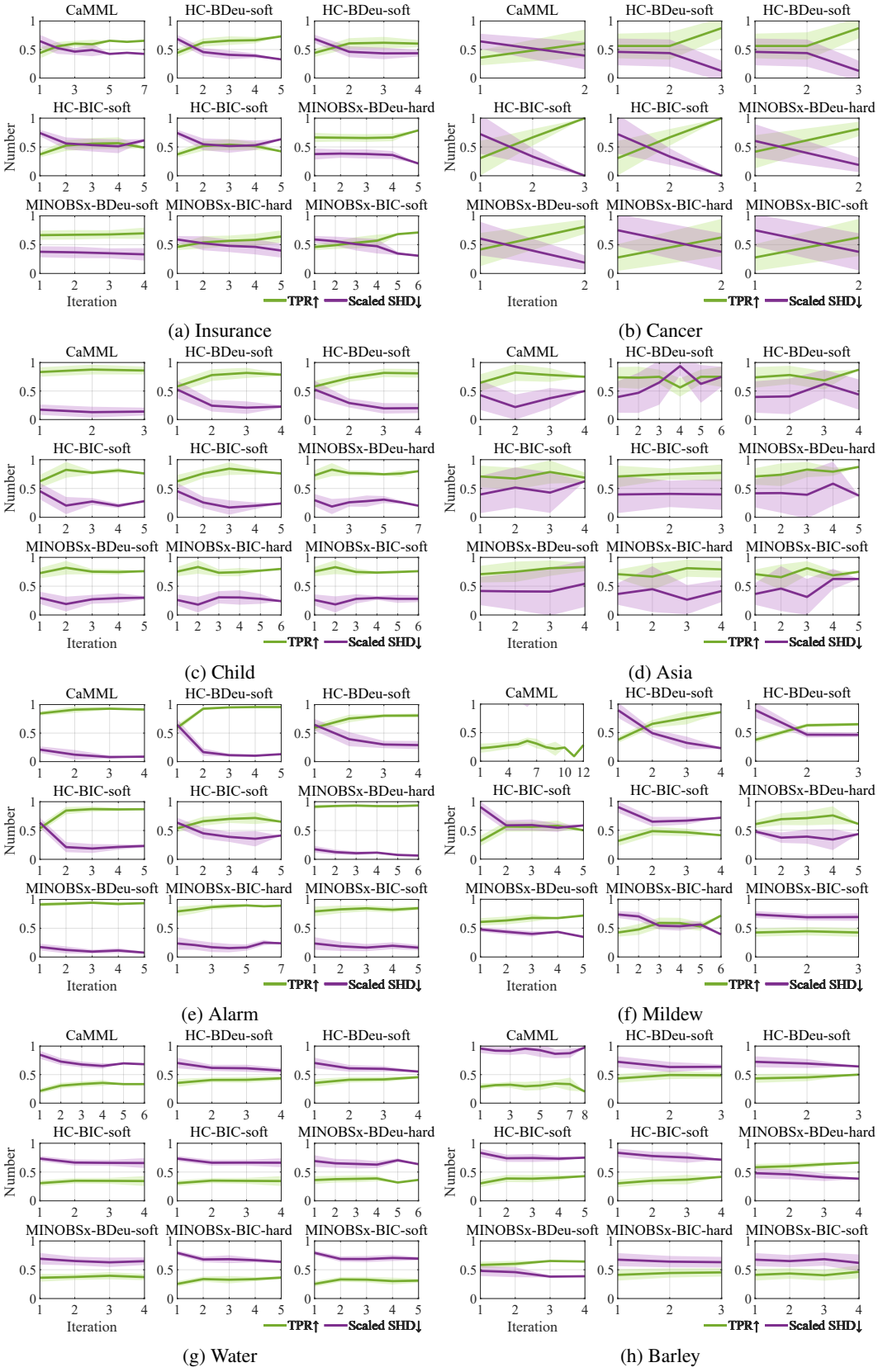
A.6.3 TREND OF CONSTRAINTS DERIVED FROM LLM OVER ITERATIONS

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 6. The setup of the reported cases remains consistent with that in Appendix A.6.2.

Observations: 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 A.6.2. 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) Insurance (b) Cancer

(c) Child (d) Asia

(e) Alarm (f) Mildew

(g) Water (h) Barley

Figure 5: TPR↑ (green line) and scaled SHD↓ (purple line) alongwith derivations (colored area) in ILS-CSL with various algorithms on various datasets.

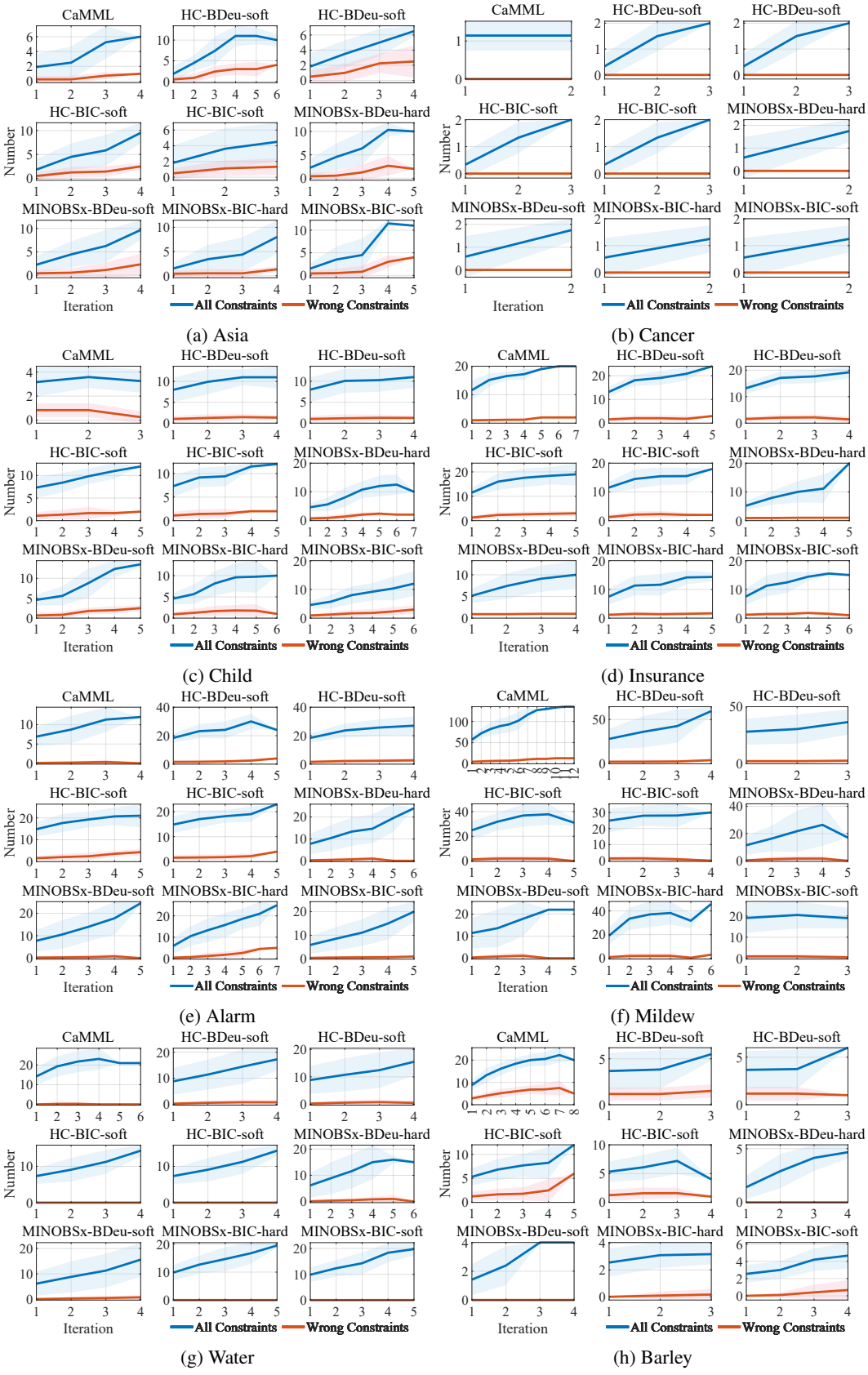


Figure 6: Number of total (blue line above) and erroneous (red line below) prior constraints along with derivations (colored area) in ILS-CSL with various algorithms on various datasets.