

Causal Modeling of Adverse Pregnancy Outcomes via Adaptive LLM Proposals

Kavimayil P. Komarasamy^{†*}, Saurabh Mathur^{‡*}, Ameet Soni[§], David M. Haas[¶],
Kristian Kersting^{‡||**}, Sriraam Natarajan[†]

[†]The University of Texas at Dallas

[‡]Technical University of Darmstadt

[§]Swarthmore College

[¶]Indiana University School of Medicine

^{||}Hessian Center for Artificial Intelligence (hessian.ai)

^{**}German Research Center for AI (DFKI)

Abstract—Adverse Pregnancy Outcomes (APOs) such as preterm birth and gestational diabetes can have long-term consequences for both the mother and child, yet an understanding of their causes remains elusive. Causal discovery in this domain is especially challenging due to a paucity of data and incomplete domain knowledge. As a result, pure data-driven methods fail, and Large Language Model (LLM) outputs remain inconsistent or contradictory. We introduce a neurosymbolic framework for generating plausible causal hypotheses that iteratively combines the broad prior knowledge of LLMs with empirical scoring on data. Our method treats the LLM as an adaptive proposal distribution, generating hypotheses that are scored against empirical data; the resulting high-scoring graphs are then used to update the LLM’s context, steering subsequent generations toward more promising regions of the hypothesis space. We evaluate our approach on a real-world clinical dataset for modeling APOs and their risk factors, comparing our results against an expert-constructed causal graph. Our method recovers all expert-validated edges and identifies additional plausible causal relations not previously listed by experts, potentially providing new insights for targeted interventions.

Index Terms—Causal Discovery, Large Language Models, Estimation of Distribution

I. INTRODUCTION

Reliable clinical reasoning requires robust causal models [1], [2], yet these models are rarely available in complex domains such as obstetrics. Modeling Adverse Pregnancy Outcomes (APOs) such as preeclampsia, gestational diabetes, and preterm birth is particularly challenging because they arise from an interplay of diverse risk factors, including maternal demographics, family history, pre-existing conditions, and lifestyle factors. While current medical literature provides high-precision knowledge of isolated links, it remains incomplete regarding the overall causal dynamics underlying APOs [3]. Moreover, data-driven causal discovery is limited by the paucity of high-quality obstetric data and the extreme difficulty of performing interventional studies in pregnant populations.

On the other hand, while Large Language Models (LLMs) can suggest causal links based on their vast training corpora, they cannot reason causally; they fail to distinguish genuine

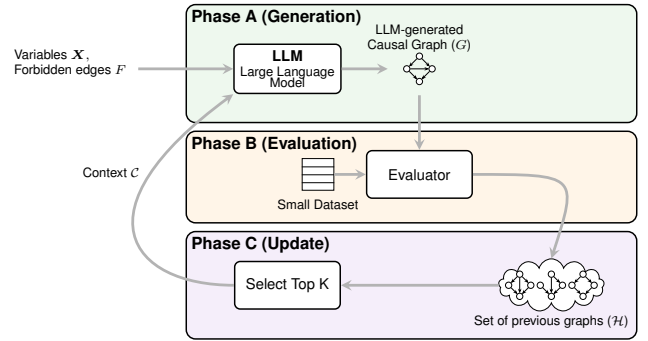


Fig. 1: **The CLARA Framework.** Our framework addresses causal discovery in data-sparse domains via a Generate-Evaluate-Update loop. It uses a pretrained LLM to *generate* diverse causal hypotheses, which are *evaluated* using a small clinical dataset. These scored graphs are used to identify patterns that consistently appear in top-scoring graphs and *update* the sampling distribution to shift mass towards more promising regions.

causation from mere association [4], [5]. Further, their outputs are highly sensitive to prompts and can result in contradictory graphs that lack grounding in empirical evidence. Existing hybrid approaches attempt to mitigate this through the theory refinement paradigm [6], [7], where an initial hypothesis is refined via local edits to improve its fit against the data. However, these methods are sensitive to their starting points, often becoming trapped in local optima that depend entirely on the quality of the initial hypothesis.

In this work, we propose CLARA (Causal Learning via Adaptive Resampling and Aggregation), a neurosymbolic framework that treats the LLM as an adaptive proposal distribution, iteratively guided by data-driven evaluation. Inspired by the MIMIC stochastic optimization algorithm [8], CLARA iteratively samples causal structures from the LLM, scores them against empirical data, and updates the prompt for subsequent generations with the highest-scoring causal structures. This process effectively narrows the LLM’s sampling mass toward high-confidence, data-validated structures.

*Equal contribution.

Specifically, we make the following key contributions: (1) We introduce CLARA, a Neurosymbolic framework that combines approximate knowledge from an LLM with incomplete expert knowledge and a small empirical dataset to generate plausible causal hypotheses. (2) We demonstrate the efficacy of our method on a real-world clinical dataset for Adverse Pregnancy Outcomes. CLARA recovers all of the edges from an expert-constructed graph and uncovers additional, plausible causal relations beyond the expert graph. (3) We show how CLARA can be made more efficient without sacrificing performance.

The rest of the paper is organized as follows: after reviewing the necessary background and related work, we outline our CLARA algorithm and analyze it. We then present empirical evidence of this algorithm on the real problem of modeling adverse pregnancy outcomes before concluding the paper by outlining the areas of future research.

II. BACKGROUND

A. Causal Bayesian Networks

Causal Bayesian Networks (CBNs) are a class of causal models (CMs) and Probabilistic Graphical Models (PGMs) [9], [10]. They represent causal relationships among a set of variables, $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ using a causal graph $G = (\mathbf{X}, E)$, where each edge $X_i \rightarrow X_j \in E$ indicates that X_i is a direct cause of X_j . This differentiates a CBN from standard BNs that only encode probabilistic dependencies; CBNs encode causal relations and support interventional reasoning, which is critical for medicine. Formally, a CBN over $\mathbf{X}$ is defined as $\langle G, \mathbf{P} \rangle$, where G is the causal Directed Acyclic Graph (Causal DAG) and $\mathbf{P} = \{P_1, \dots, P_n\}$ is the set of local conditionals over each variable given its parents. The CBN factorizes the joint distribution over an assignment $\mathbf{X} = \mathbf{x}$ as the product of local conditionals: $P(\mathbf{x}) = \prod_i P_i(X_i \mid \text{Pa}_i(\mathbf{x}))$.

The focus of our work is generating plausible causal hypotheses: the task of identifying the graph G that accurately represents the causal relations among variables $\mathbf{X}$. Data-driven methods such as Peter-Clark (PC [11]), Fast Causal Inference (FCI [12]), and Greedy Equivalence Search (GES [13]) learn the causal graph from observational data. They establish causation by combining patterns induced from large amounts of data with structural assumptions such as faithfulness, causal Markov condition, and causal sufficiency (see Section IV-A for variable selection details). However, clinical datasets are typically small and noisy, with many relevant variables only weakly observed [14]. Moreover, learning the optimal BN structure from data is difficult because the number of possible graph structures grows super-exponentially with the number of variables [15]. As a result, algorithms that rely mainly on statistical signals such as conditional independence may fail to recover plausible causal structures in complex and data-scarce domains.

B. LLM-augmented Theory Refinement

Theory refinement [16] is a class of hybrid methods that address limitations of purely data-driven learning by ex-

ploiting domain knowledge. The use of domain knowledge allows theory refinement to reduce the search over a super-exponential number of possible graph structures to a local search in the neighborhood of an initial domain knowledge-based structure. This baseline BN structure is constructed using incomplete domain knowledge obtained from experts and is refined by local hill-climbing search; each step applies one of 3 types of atomic editing operations: adding, deleting, or reversing directed edges. Each operation is selected to maximize an empirical structure score, such as the Bayesian-Dirichlet (BD [17]) score and the Bayesian Information Criterion (BIC [18], [19]). However, as the number of variables and potential relationships increases, specifying a good initial graph becomes challenging.

Recent work has addressed this issue using pretrained Large Language Models (LLMs [20]). These models have emerged as effective sources of *approximate domain knowledge* that can be used to create the initial hypothesis for theory refinement [7], [21]. These models capture causal domain knowledge from their vast training corpora, which include academic literature. However, while LLMs can generate fluent text, their reasoning capabilities remain limited, especially in causal inference [4]. Their outputs are stochastic and inconsistent across different runs. Since hill-climbing is a local search, relying on a single hypothesis from an LLM response may result in a structure that misses clinical nuances.

C. Estimation of Distribution Algorithms

To overcome the sensitivity of theory refinement to poor initial hypotheses, we use Estimation of Distribution Algorithms (EDAs) [22]. Unlike traditional optimization methods that track a single solution at a time, EDAs are a class of stochastic optimization methods that maintain and iteratively update a distribution over the entire search space. This population-based approach allows the algorithm to explore multiple neighborhoods of the search space simultaneously, reducing the risk of being trapped in local optima.

We specifically adopt the framework of Mutual-Information-Maximizing Input Clustering (MIMIC) [8], which provides a principled mechanism for optimization by transforming a simple initial distribution into one concentrated around optimal solutions through a Sample-Evaluate-Update loop. MIMIC draws samples from the distribution, evaluates them against the objective function, and updates the sampling distribution by learning the common structural features of the top-scoring samples. This process is repeated, refining the sampling distribution until it concentrates around a set of high-scoring candidate solutions.

CLARA instantiates this loop with two key substitutions. Rather than learning an explicit distribution from the top-scoring samples and drawing new candidates from it, CLARA uses an LLM as the sampling distribution, and updates subsequent generations by replacing its in-context examples with the top-scoring samples. This yields the Generate-Evaluate-Update loop described in Section III.

III. CAUSAL LEARNING VIA ADAPTIVE RESAMPLING AND AGGREGATION

We aim to build a causal model by integrating approximate knowledge from an LLM, incomplete expert knowledge, and empirical evidence from data. We formalize this as the following problem:

Given: A dataset $\mathcal{D}$ over discrete variables $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$, incomplete expert knowledge in the form of a set of forbidden edges $\mathbf{F}$, and a pretrained Large Language Model $\mathcal{O}$.

To Do: Find a causal graph G that accurately captures the causal relations over $\mathbf{X}$ by reconciling these knowledge sources.

The primary challenge in this domain is the inherent incompleteness of information sources. While clinical experts provide high-precision knowledge, their expertise is often siloed within specific sub-specialties, leaving significant blind spots in the full causal structure. Further, the paucity of observational data prevents purely data-driven algorithms from fully capturing the causal structure, yielding overly sparse graphs. Finally, while Large Language Models (LLMs) act as broad knowledge sources, their outputs are fundamentally stochastic and unreliable; they typically fail to distinguish genuine causation from association and are highly sensitive to initial prompting, leading to unstable and inaccurate performance in one-shot causal graph generation.

To bridge this gap, we introduce Causal Learning via Adaptive Resampling and Aggregation (CLARA). As outlined in Fig. 1, CLARA instantiates a neurosymbolic system (Type 2: Symbolic[Neuro] [23]) that finds causal graphs through a Sample-Evaluate-Update loop, gradually shaping the broad prior of a pretrained LLM to a distribution concentrated around data-validated plausible causal graphs.

A. Sampling Causal Graphs

CLARA starts by querying the LLM $\mathcal{O}$ for a batch of K candidate causal hypotheses. This jumpstarts the search using the LLM’s broad domain knowledge while mitigating the instability of committing to a single output. Each sample is obtained by prompting the LLM with the list of variables $\mathbf{X}$, along with brief descriptions, and expert knowledge in the form of forbidden edges $\mathbf{F}$. Each candidate graph sampled from the LLM that violates forbidden edge and acyclicity constraints is discarded and resampled.

B. Evaluating the Causal Graphs using Structure Score

CLARA scores each LLM proposal G' against the dataset $\mathcal{D}$ using the Bayesian Information Criterion (BIC) [18] as a structure score. BIC balances the model’s goodness of fit to the data against its complexity, and is defined as

$$\text{BIC}(G') = \log P(\mathcal{D} | G') - \frac{\log N}{2} \sum_{i=1}^n q_i (|X_i| - 1) \quad (1)$$

where $\log P(\mathcal{D} | G')$ is the log-likelihood of the data $\mathcal{D}$ given graph G' , N is the number of observations, $q_i = \prod_{X_j \in \text{Pa}(X_i)} |X_j|$ is the number of parent configurations of X_i , and $|X_i|$ is the number of states of X_i . In discrete Bayesian networks, each variable is parameterized by a conditional probability table (CPT) over all configurations of its parent variables, and the complexity penalty scales with the number of parameters. The complexity penalty term in BIC can massively penalize graph structures with a large number of parents and parameters, causing the score to underfit the data by favoring overly sparse graphs and pruning meaningful dependencies [24]. To mitigate this, we account for local structure in the form of Context-Specific Independencies (CSI) [25] by representing each local conditional as a decision tree [26] instead of a full CPT. Each leaf stores a conditional distribution over the child variable conditioned on the parent assignments along the corresponding root-to-leaf path. This representation merges parent configurations when further splitting does not reduce the description length under the minimum description length criterion [19]. The resulting *Tree BIC* score replaces the product of parent cardinalities in the Standard BIC penalty with the number of decision-tree leaves,

$$\text{TreeBIC}(G') = \log P(\mathcal{D} | G') - \frac{\log N}{2} \sum_{i=1}^n l(T_i) (|X_i| - 1) \quad (2)$$

where $l(T_i)$ is the number of leaves in the decision tree, yielding a smaller penalty for context-specific structure. The scored graphs are then added to the history set $\mathcal{H}$ and ranked by their Tree BIC scores, ensuring that only the top-scoring graphs contribute to the context for the next generation.

C. Updating the Sampling Distribution

Finally, to update the LLM sampling distribution for the next round of generation, CLARA constructs an in-context representation $\mathcal{C}$ from the best-scoring graphs in $\mathcal{H}$ to condition the next LLM generation, effectively narrowing the sampling distribution toward high-confidence structures. Since CLARA prompts the LLM to output graphs as lists of directed edges, we use a compatible representation (adjacency lists) in the in-context examples to ensure consistent semantics between prompt and output. As an alternative to constructing $\mathcal{C}$ from the full high-scoring graphs, we consider a compressed representation consisting of the edges common to all high-scoring graphs. We refer to this as the common edges representation and evaluate it as a more efficient alternative in Section IV. This completes CLARA’s generate-evaluate-update loop, ensuring that each generation is conditioned on structures supported by the LLM’s broad prior and empirical data.

D. The CLARA Algorithm

Algorithm 1 presents the CLARA algorithm. It finds plausible causal structures by adapting the MIMIC algorithm’s Generate-Evaluate-Update loop. Traditional MIMIC [8] iteratively learns a simple density estimator such as a chain-structured Bayesian network over a bit-string encoding of the

Algorithm 1: Causal Learning via Adaptive Resampling and Aggregation (CLARA)

Input: Variables $\mathbf{X}$, Observational Data $\mathcal{D}$, Forbidden Edges $\mathbf{F}$, Large Language Model $\mathcal{O}$, Batch Size K , Iterations T

Output: Final causal graph G_{Final}

```
 $\mathcal{H} \leftarrow \emptyset$            // Initialize history set
 $\mathcal{C}_0 \leftarrow \emptyset$  // Initialize representation set
while  $|\mathcal{H}| < K$  // Generate initial graphs
do
   $G' \leftarrow \text{SAMPLEFROMLLM}(\mathcal{O}, \mathbf{X}, \mathbf{F}, \mathcal{C}_0)$ 
   $G \leftarrow \text{EVALUATEWITHDATA}(G, \mathcal{D}, \mathbf{F})$ 
   $\mathcal{H} \leftarrow \mathcal{H} \cup \{G\}$ 
for  $t = 1$  to  $T$  do
   $G_{\text{TopK}} \leftarrow \text{SELECTTOPK}(\mathcal{H}, K)$  // Top  $K$ 
  graphs ranked by Tree BIC
   $\mathcal{C}_t \leftarrow \text{INDUCEREPRESENTATION}(G_{\text{TopK}})$ 
  // Full Graphs or common edges
   $G' \leftarrow \text{SAMPLEFROMLLM}(\mathcal{O}, \mathbf{X}, \mathbf{F}, \mathcal{C}_t)$ 
   $G \leftarrow \text{EVALUATEWITHDATA}(G, \mathcal{D}, \mathbf{F})$ 
   $\mathcal{H} \leftarrow \mathcal{H} \cup \{G\}$  // Update history
 $G_{\text{TopK}} \leftarrow \text{SELECTTOPK}(\mathcal{H}, K)$ 
return AGGREGATECANDIDATES( $G_{\text{TopK}}$ )
// Return union of best candidates
```

solution space and samples new candidates from the learned distribution. In contrast, CLARA replaces this learning-and-sampling step with a pretrained LLM, leveraging its encoded domain knowledge while conditioning generation on an in-context representation $\mathcal{C}$ constructed from the top-scoring graphs. The algorithm initializes the history set $\mathcal{H}$ by sampling an initial batch of K causal graphs using `SAMPLEFROMLLM`, without any in-context examples ($\mathcal{C}_0 = \emptyset$). This initial set of graphs is scored using `EVALUATEWITHDATA`, which computes the `TreeBIC` score. The evaluated graphs and their scores are then stored in $\mathcal{H}$.

By treating the LLM as an adaptive proposal distribution, CLARA implements the MIMIC Generate-Evaluate-Update optimization loop. At each iteration, `SELECTTOPK` identifies the K highest-scoring graphs stored in $\mathcal{H}$, `INDUCEREPRESENTATION` updates $\mathcal{C}$ based on the current set of top- K causal graphs, and `SAMPLEFROMLLM` generates a new candidate causal graph conditioned on this context. CLARA evaluates each new LLM proposal using `EVALUATEWITHDATA`, which computes the structure score of the proposed causal graph based on $\mathcal{D}$. The evaluated graph and its score are then appended to the history set $\mathcal{H}$ before influencing subsequent generations, rather than treating the raw LLM output as a final hypothesis. While the construction of an in-context representation $\mathcal{C}$ serves as a distribution update, the scoring process ensures that only high-scoring candidates are used to construct the next generation context. Finally, once all

T iterations are complete, the top- K high-scoring graphs in $\mathcal{H}$ are aggregated using a union-based strategy to construct the final causal graph. The details of the aggregation and cycle-breaking procedure are described in Section IV-B. This Generate-Evaluate-Update loop allows CLARA to effectively address the one-shot limitation of standard LLM-based causal graph generation with a global optimization strategy, which we evaluate empirically in Section IV-C.

IV. EMPIRICAL EVALUATION

We aim to answer the following questions:

Q1: Does CLARA’s combination of LLM priors and search yield better causal hypotheses than purely LLM-based generation, pure structure search, and LLM-initialized local search?

Q2: Does CLARA yield better causal hypotheses than data-driven causal discovery methods in data-scarce and noisy domains?

Q3: Does replacing DAGs with structural summaries reduce prompt size while maintaining performance?

Q4: Does CLARA propose plausible causal hypotheses in the real obstetrics domain, closely matching domain knowledge?

A. Dataset Description

To answer these questions, we employ two datasets. First, we generated a synthetic dataset based on the ALARM benchmark, a Bayesian network originally developed for patient monitoring and alarm systems [27]. The network contains 37 nodes and 46 edges, and both the ground-truth causal graph and conditional probability tables (CPTs) are obtained from the bnlearn repository¹. We sampled 3,000 observations from this network based on the provided probability tables and introduced 20% entry-wise categorical noise, where each variable value is independently replaced with a randomly selected valid state with probability 0.2 to simulate imperfect observational conditions.

Second, we used real clinical data from the nuMoM2b study (Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be) [3], a large-scale longitudinal study of first-time mothers with singleton pregnancies. We focused on variables collected during the first prenatal visit to model the causal impact of early-pregnancy risk factors on APOs. The set of risk factors comprises 9 variables across four clinically salient domains: **demographics, family history, pre-existing conditions, and lifestyle factors**. We consider 4 adverse pregnancy outcomes: **preeclampsia, new hypertension, gestational diabetes, and preterm birth** (both spontaneous and medically indicated). These variables were selected by our clinical experts to ensure causal sufficiency to the best of current clinical knowledge. To ensure the integrity of the causal hypotheses generation, we performed targeted preprocessing. We excluded data from subjects with pre-existing diabetes to

¹<https://www.bnlearn.com/bnrepository/>

TABLE I: **Baseline characteristics and pregnancy outcomes.** Variables and their frequencies in our dataset. Row colors indicate variable type: demographic, family history, pre-existing conditions, lifestyle factors, pregnancy outcomes. All risk factors were measured at the start of pregnancy.

Variable	%
Race	
Non-Hispanic Asian	4.2
Non-Hispanic Black	11.2
Non-Hispanic White	67.8
Hispanic	12.3
Other	4.4
Age	
≤ 21	15.9
21–35	77.0
> 35	7.0
BMI	
≤ 18	1.3
18–25	54.1
> 25	44.6
Family history of hypertension (HTNHist)	45.4
Family history of type 2 diabetes (DiabHist)	20.7
High blood pressure (HiBP)	2.7
Polycystic ovary syndrome (PCOS)	4.8
Smoked tobacco in 3 months before pregnancy (Smoked)	14.9
Adequate physical activity (PhysAc)	66.1
Preeclampsia (PreEc)	5.9
New hypertension (NewHTN)	17.7
Gestational diabetes (GDM)	3.8
Preterm birth (PTB)	7.7

isolate the mechanisms associated with gestational diabetes. Further, we excluded all data points containing missing values to ensure the structural search relied strictly on observed clinical signals. Following these steps, we retained a final dataset of 3,856 examples. The full list of variables and their frequency distributions is presented in Table I.

B. Method and Baselines

We compare CLARA with three types of baselines: (i) **purely data-driven discovery**, (ii) **one-shot LLM generation**, and (iii) **LLM-augmented theory refinement**, as well as MIMIC [8] as a non-LLM EDA ablation isolating CLARA’s iterative search loop from its LLM prior. The data-driven baselines are the PC [11] and FCI [12] algorithms to represent standard approaches for identifying causal structures from observational data. Both data-driven algorithms use the chi-squared conditional independence test with a significance level $\alpha = 0.05$, consistent with the discrete-valued variables in both the ALARM and nuMoM2b datasets. LLM-dependent methods use GPT-5.2 [28] and Llama-3.3-70B-Instruct([29]) as the pretrained generators. Temperature is set to 0.7 across all LLM-based methods. The maximum token size is set to 8192 for Llama-3.3-70B-Instruct (other parameters default for both models). To reduce sampling variance, each candidate graph in CLARA is formed by the union of five independently sampled responses. Cycles introduced during aggregation are

broken by removing the edge with the lowest frequency among five responses, with ties broken lexicographically (in ALARM, lexicographic resolution is used in 26% of cycle-breaking events whereas no cycles occur in nuMoM2b). The forbidden edge set **F** encoding temporal order constraints as incomplete expert knowledge is available to all methods for nuMoM2b, whereas ALARM has no blacklist edges.

In addition, we evaluate two variants of CLARA to assess the impact of different components of the system: (1) CLARA (as fully implemented), using full graphs ranked by BIC score; (2) CLARA (Common edges), using representations enforcing direct adjacencies, included in-context in the prompt during the LLM-guided exploration. A common edge is included in the structural summary representation only if it appears in all top-K scoring graphs to reflect full agreement across the top-K set. For all experiments, we set the batch size $K=7$ and the number of iterations $T=5$, as this configuration achieves the lowest Structural Intervention Distance (SID) [30] in the sensitivity analysis (Table III), with no additional improvements observed from further iterations. Since LLM-based methods are inherently stochastic, all LLM-based methods were run five times independently, and results are reported as mean $\pm$ standard deviation across runs. Deterministic methods (PC, FCI) are reported as single values.

C. Evaluation Metrics

To evaluate structural correctness, we compare the learned causal graphs against a reference graph constructed by our obstetrics expert based on current medical consensus (Figure 2). As obstetrics remains an active area of research, the expert-constructed causal graph in Figure 2 encodes all the causal relationships known to the best of our clinicians’ knowledge. While we report **Structural Hamming Distance (SHD)** [31], **Recall**, and **Precision** to measure topological alignment with this graph, we primarily focus on **Structural Intervention Distance (SID)** [30]. In high-stakes clinical domains like obstetrics, the functional reliability of a model in reasoning about the adjustment set for an intervention is more critical than exact adjacency recovery. We note that SHD treats every edge addition, deletion, and reversal as an equally severe error, regardless of whether that edge affects the adjustment set needed for a correct intervention.

A graph can still have higher SHD while supporting correct interventional reasoning. *For this reason, we treat SHD as a secondary metric and SID as our primary metric to support our comparative claims in discussion.* We further analyze the additional edges discovered beyond the expert reference based on whether they represent potential clinical discoveries.

D. Results and Discussion

(Answer, Q1:) To answer Q1, we compared the causal networks generated by CLARA and the baselines to the ground-truth network in the synthetic ALARM dataset and an expert-constructed network in the real-world numom2b dataset. Table II quantifies the divergence of the networks generated by each method with the reference network for each domain in

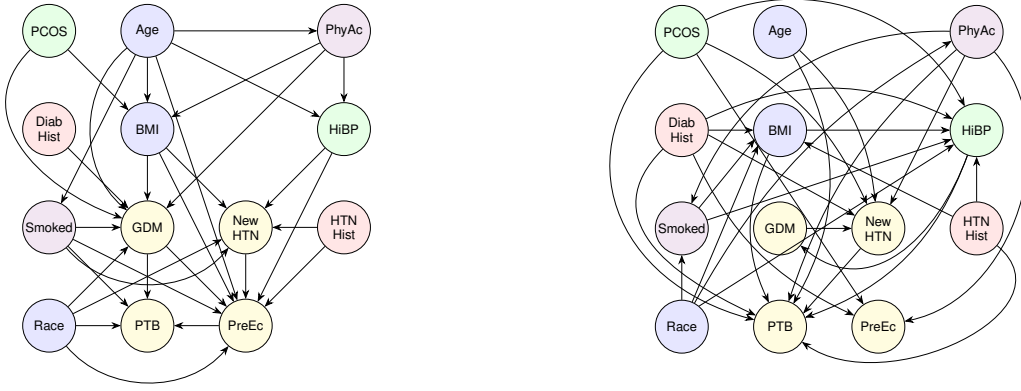


Fig. 2: **Expert-elicited causal graph** (left) and **additional edges discovered by CLARA** (right) for the obstetrics domain. Node color indicates variable type: **demographic**, **family history**, **pre-existing conditions**, **lifestyle factors**, **pregnancy outcomes**.

TABLE II: **Comparative evaluation of CLARA and baselines on ALARM and nuMoM2b domains.**

Method	ALARM				nuMoM2b			
	SHD ↓	SID ↓	Recall ↑	Precision ↑	SHD ↓	SID ↓	Recall ↑	Precision ↑
PC	35	314	0.5	0.6	34	83	0.1	0.3
FCI	46	368	0.1	0.5	32	79	0.0	0.0
MIMIC	66.2±10.4	473.0±15.4	0.0	0.1	37.6±3.1	73.6±2.5	0.1	0.2
LLM: GPT-5.2								
LLM (One-shot)	48.2±2.7	205.7±18.0	0.6	0.4	29.2±2.7	8.8±4.4	1.0	0.5
LLM + Theory Refinement	41.6±3.7	194.2±12.6	0.5	0.5	27.4±1.7	21.0±0.0	0.8	0.5
CLARA	66.0±4.1	139.8±19.8	0.7	0.3	28.4±1.2	0.0±0.0	1.0	0.5
CLARA (Common edges)	67.8±5.5	153.6±24.0	0.7	0.3	29.0±2.4	0.0±0.0	1.0	0.5
LLM: Llama-3.3-70B-Instruct								
LLM (One-shot)	74.4±2.0	322.0±33.0	0.3	0.2	26.8±3.1	47.4±3.7	0.4	0.5
LLM + Theory Refinement	66.6±2.1	321.2±27.6	0.3	0.2	26.4±2.8	51.2±4.1	0.4	0.5
CLARA	108.8±4.0	245.0±30.3	0.4	0.1	26.4±1.3	31.4±3.3	0.6	0.5
CLARA (Common edges)	110.8±5.9	247.4±23.8	0.4	0.1	25.8±1.4	29.6±4.8	0.6	0.5

terms of differences in structure (SHD, Precision, Recall) and differences in causal conclusions (SID).

On the synthetic ALARM dataset, CLARA achieves the lowest SID across all methods, 32% better than one-shot LLM generation, 55% better than data-driven causal discovery, and 28% better than the hybrid LLM-initialized theory refinement. This trend holds for both LLMs (GPT-5.2 and Llama-3.3-70B-Instruct), indicating that CLARA’s search procedure contributes to the improvement independent of the underlying LLM, though the magnitude of improvement varies by model. While LLM-generated networks captured more causal dependencies than purely data-driven causal discovery, they also captured more spurious relationships, increasing their recall at the cost of precision. Additionally, while LLM-initialized theory refinement eliminated some spurious edges from the LLM-generated network, its greedy search also eliminated weaker causal edges, increasing precision at the cost of recall. However, this tradeoff is unsuitable for high-stakes domains like medicine, where assuming causal independence is more costly than assuming causal dependence. CLARA navigates this tradeoff more cautiously, capturing more causal dependencies

by iteratively combining the LLM’s broad prior with data-driven evaluation. Unlike theory refinement’s greedy, one-edge-at-a-time pruning, CLARA’s repeated resampling can retain edges that are individually weak but become informative jointly with other parents.

The same pattern holds on the real-world nuMoM2b dataset (Table II). CLARA achieves the lowest SID across all methods, outperforming one-shot LLM generation, data-driven causal discovery, and the hybrid LLM-initialized theory refinement across both LLMs. Moreover, in contrast to the synthetic ALARM domain, theory refinement eliminates a lot more causal edges, increasing the number of causal reasoning errors by 138% relative to LLM One-shot. Providing the same forbidden edge set $\mathbf{F}$ to the data-driven baselines on nuMoM2b did not alter their scores, suggesting that CLARA’s improvements stem primarily from the LLM-derived priors. Consistent with this, MIMIC, which performs the same iterative resampling procedure without an LLM-informed prior, only marginally outperforms the purely data-driven baselines on nuMoM2b and yields the worst SID of any method on ALARM despite sharing CLARA’s architecture.

Overall, CLARA generates causal networks whose causal conclusions match the reference networks at a higher rate than purely LLM-based, data-driven causal discovery, and hybrid LLM-initialized theory refinement baselines. Therefore, we can answer **Q1** affirmatively.

(Answer, Q2:) To answer Q2, we analyze Figure 3, which plots SID across dataset sizes and noise levels for PC, FCI, and CLARA on the synthetic ALARM domain. CLARA maintains a low and stable SID across all combinations of dataset size and noise level. Across noise levels, CLARA’s SID remains largely unchanged as the proportion of randomly perturbed entries increases from 20% to 60%. Across dataset sizes, CLARA’s SID remains similarly low even as the number of samples is reduced, showing little dependence on the amount of available data. In contrast, while PC narrows the gap with CLARA as data increases at low noise (20%), this improvement disappears at higher noise levels (40% and 60%), where its SID becomes non-monotonic and degrades with an increase in dataset size at the highest noise level. FCI remains elevated and non-monotonic across all combinations of dataset size and noise level, showing no clear benefit from additional data at any noise level. Notably, PC and FCI fail to match CLARA’s SID even when the dataset size is tripled. This stability across both axes indicates that CLARA generates better causal hypotheses than purely data-driven methods in both data-scarce and noisy clinical settings. We can therefore answer **Q2** affirmatively.

(Answer, Q3:) To answer Q3, we examine the performance versus prompt size tradeoffs by comparing CLARA with a variant that replaces the full list of top-scoring graphs in the prompt with structural summaries in the form of edges common to top-scoring graphs. This yields up to three times fewer prompt tokens as shown in Table IV. This variant achieves an SID closely matching CLARA on three of the four dataset–LLM combinations. On ALARM with GPT-5.2, SID of the common edges variant is worse than CLARA by 10%, but it is still 21% better than the other baselines. Overall, these results indicate that replacing full DAGs with structural summaries reduces prompt size while maintaining performance close to that of fully implemented CLARA. We can therefore answer Q3 affirmatively.

(Answer, Q4:) To answer Q4, we compare the edges discovered by CLARA with the expert-constructed graph. CLARA recovers all 31 expert-validated causal edges in Figure 2 and proposes 30 additional edges. To assess their validity, we computed a positive likelihood ratio (LR+) for each using data and presented them to our obstetric clinical expert for rating against current medical consensus. The clinician rated 26 (87%) as established, and the remaining 4 (13%) as plausible: PCOS→PTB, HiBP→GDM, DiabHist→PTB (likely mediated rather than direct), and Race→PhyAc (varying in strength by Race). We identify three patterns in the additional edges.

First, CLARA identifies causal relationships among risk factors themselves. CLARA proposes family history of diabetes and hypertension as contributing causes of pre-pregnancy BMI and high blood pressure, consistent with evidence from a UK population-based cohort [32] and a Japanese longitudinal

study [33]. CLARA further proposes race as a contributing cause of BMI [34], physical activity [35], and smoking [36], relationships supported by US-based evidence linking racial disparities to socioeconomic determinants of lifestyle. CLARA also proposes PCOS as a contributing cause of high blood pressure, confirmed by a meta-analysis of 30 studies in reproductive-age women [37].

Second, several additional edges are direct connections whose literature-established relationships are typically mediated by variables absent from the dataset. The direct edges from maternal age [38] and pre-pregnancy BMI [39] to preterm birth are well-supported but partly explained by intermediate factors such as gestational diabetes or preeclampsia. CLARA’s prior knowledge recovers these reduced-form direct effects in the absence of the mediating variables. Similarly, the direct edges from PCOS to preeclampsia and preterm birth are consistent with a meta-analysis showing a threefold increased preeclampsia risk in women with PCOS [40].

Third, CLARA identifies a sequential complication cascade among APOs: gestational diabetes as a contributing cause of new hypertension, supported by evidence that GDM substantially increases the risk of hypertensive disorders of pregnancy through shared metabolic mechanisms [41], and new hypertension as a contributing cause of preterm birth, consistent with evidence from a large Chinese prospective cohort [42].

These findings, validated by our obstetric clinical expert, indicate that CLARA proposes plausible causal hypotheses closely matching domain knowledge in obstetrics. We can therefore answer Q4 affirmatively.

V. CONCLUSION

We addressed the task of generating plausible causal hypotheses for Adverse Pregnancy Outcomes (APOs). To do so, we introduced CLARA, a neurosymbolic framework that reconciles the broad priors encoded in pretrained Large Language Models with the empirical rigor of data-driven structure scoring. By framing the causal discovery process as a Sample-Evaluate-Update loop, CLARA effectively mitigates the stochasticity and unreliability inherent in one-shot neural proposals while overcoming the limitations of data sparsity in complex clinical domains. Moving forward, there are several promising directions to enhance this paradigm. First, the framework could be extended by incorporating more expressive constraint classes, such as qualitative influence statements indicating whether a cause suppresses or reinforces its effect, and constraints capturing time-varying and other context-specific clinical dependencies. Second, to address the computational challenges of high-dimensional domains, a hierarchical version of the estimation of distribution algorithm could be developed to decompose the causal search space into manageable sub-networks. Third, a systematic empirical comparison of the current union-based aggregation strategy with alternative approaches, such as higher-order DAG aggregation methods, would provide a better understanding of the trade-offs between different aggregation schemes. Finally, exploring methods for fine-tuning the underlying language models to

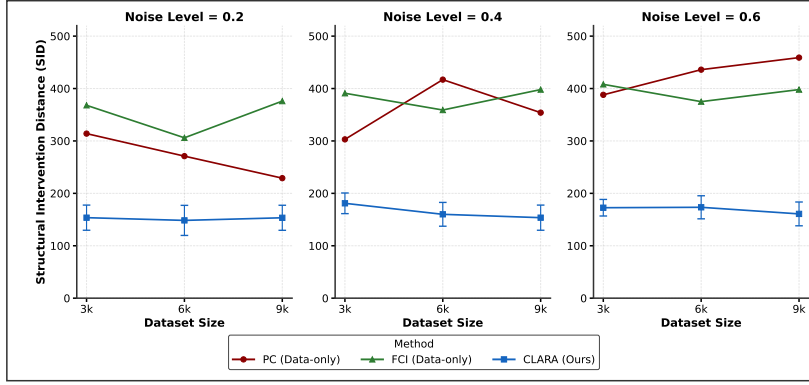


Fig. 3: **Sensitivity of Structural Intervention Distance (SID) to noise and data size.** SID of PC, FCI, and CLARA across dataset sizes (3k, 6k, 9k) at noise levels 0.2, 0.4, 0.6. Each panel corresponds to a fixed noise level. PC and FCI are deterministic and reported as single-run values. CLARA SID values are the mean $\pm$ standard deviation across five independent runs (shown as error bars) using GPT-5.2 with $K=7$, $T=5$. Lower SID indicates fewer errors in predicted interventional distributions, i.e., better recovery of causal structure.

TABLE III: **Hyperparameter sensitivity analysis.** Mean and standard deviation of SID scores are reported for combinations of K (top- K best-scoring graphs) and T (number of iterations) on the ALARM dataset using GPT-5.2 as proposal LLM. Lower SID is better.

K/T	5	10	15	20
3	185.4 $\pm$ 25.9	185.4 $\pm$ 25.9	185.4 $\pm$ 25.9	185.4 $\pm$ 25.9
5	176.0 $\pm$ 28.6	176.0 $\pm$ 28.6	173.2 $\pm$ 27.7	173.2 $\pm$ 27.7
7	153.6 $\pm$ 24.0	153.6 $\pm$ 24.0	153.6 $\pm$ 24.0	153.6 $\pm$ 24.0

TABLE IV: **Average prompt token counts by CLARA variant and dataset.** Average prompt size (tokens) per variant and dataset, measured across 5 independent runs for GPT-5.2 and Llama-3.3-70B.

Dataset	LLM	Full	Common edges
ALARM	GPT-5.2	3,825	1,377
	Llama-3.3-70B-Instruct	4,164	1,364
nuMoM2b	GPT-5.2	2,535	1,195
	Llama-3.3-70B-Instruct	1,900	1,111

make them more specialized and reliable generators of causal hypotheses remains important future work. Overall, together with these directions, CLARA presents an effective framework for causal discovery in complex, data-scarce medical domains.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support from NIH awards R01HD101246 and R01NS133142, and the Cluster of Excellence “Reasonable AI” funded by the German Research Foundation (DFG) under Germany’s Excellence Strategy, EXC-3057.

REFERENCES

- [1] B. Kuipers and J. P. Kassirer, “Causal reasoning in medicine: analysis of a protocol,” *Cognitive Science*, 1984.
- [2] A. Mathur, “Teaching clinical reasoning through cause-and-effect thinking: A framework for modern medical education,” *The Clinical Teacher*, 2025.
- [3] D. M. Haas, C. B. Parker, D. A. Wing, S. Parry, W. A. Grobman, B. M. Mercer, H. N. Simhan, M. K. Hoffman, R. M. Silver, P. Wadhwa *et al.*, “A description of the methods of the nulliparous pregnancy outcomes study: monitoring mothers-to-be (numom2b),” *AJOG*, 2015.
- [4] M. Zečević, M. Willig, D. S. Dhami, and K. Kersting, “Causal parrots: Large language models may talk causality but are not causal,” *TMLR*, 2023.
- [5] S. Mathur, R. Singh, M. Skinner, P. Radivojac, D. M. Haas, L. Raman, and S. Natarajan, “Llms for causal reasoning in medicine? a call for caution,” in *Proceedings of the 13th ACM IKDD International Conference on Data Science*, 2025, pp. 164–172.
- [6] W. Buntine, “Theory refinement on bayesian networks,” in *UAI*. Elsevier, 1991.
- [7] S. Mathur, R. Singh *et al.*, “Llm-guided causal bayesian network construction for pediatric patients on ecmo,” in *AIME*. Springer, 2025.
- [8] J. De Bonet, C. Isbell, and P. Viola, “Mimic: Finding optima by estimating probability densities,” in *Advances in Neural Information Processing Systems*, M. Mozer, M. Jordan, and T. Petsche, Eds., vol. 9. MIT Press, 1996.
- [9] D. Koller and N. Friedman, *Probabilistic Graphical Models*. MIT press, 2009.
- [10] J. Pearl, *Causality*. Cambridge university press, 2009.
- [11] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.
- [12] P. Spirtes, “An anytime algorithm for causal inference,” in *AISTATS*. PMLR, 2001, pp. 278–285.
- [13] D. M. Chickering, “Optimal Structure Identification With Greedy Search,” *JMLR*, 2002.
- [14] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, “A review of challenges and opportunities in machine learning for health,” *AMIA Summits on Translational Science Proceedings*, 2020.
- [15] D. M. Chickering, D. Heckerman, and C. Meek, “Large-sample learning of bayesian networks is np-hard,” *J. Mach. Learn. Res.*, 2004.
- [16] R. J. Mooney and J. W. Shavlik, “A recap of early work on theory and knowledge refinement,” in *Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021)*, Stanford University, Palo Alto, California, USA, March 22-24, 2021, ser. CEUR Workshop Proceedings, A. Martin, K. Hinkelmann, H. Fill, A. Gerber, D. Lenat, R. Stolle, and F. van Harmelen, Eds., vol. 2846, 2021.
- [17] D. Heckerman, D. Geiger, and D. M. Chickering, “Learning bayesian networks: The combination of knowledge and statistical data,” *Machine Learning*, 1995.
- [18] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, 1978.
- [19] J. Rissanen, “Modeling by shortest data description,” *Automatica*, 1978.

- [20] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.
- [21] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, "Language models as knowledge bases?" in *EMNLP-IJCNLP*, 2019.
- [22] M. Hauschild and M. Pelikan, "An introduction and survey of estimation of distribution algorithms," *Swarm and Evolutionary Computation*, 2011.
- [23] H. Kautz, "The third ai summer: Aai robert s. engelmore memorial lecture," *Ai magazine*, 2022.
- [24] S. Natarajan, W.-K. Wong, and P. Tadepalli, "Structure refinement in first order conditional influence language," in *ICML Workshop on Open Problems in Statistical Relational Learning*, 2006.
- [25] C. Bouillier, N. Friedman, M. Goldszmidt, and D. Koller, "Context-specific independence in bayesian networks," in *UAI*, 1996.
- [26] N. Friedman and M. Goldszmidt, "Learning bayesian networks with local structure," in *UAI*, 1996.
- [27] I. A. Beinlich, H. J. Suermondt, R. M. Chavez, and G. F. Cooper, "The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks," in *AIME 89*. Springer Berlin Heidelberg, 1989.
- [28] OpenAI, "Chatgpt: GPT-5.2 language model," <https://openai.com/chatgpt>, 2025.
- [29] Meta, "Llama 3.3 70b instruct," <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>, 2024.
- [30] J. Peters and P. Bühlmann, "Structural Intervention Distance for Evaluating Causal Graphs," *Neural Computation*, vol. 27, no. 3, 2015.
- [31] J. Wahl and J. Runge, "Separation-based distance measures for causal graphs," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2025, pp. 3412–3420.
- [32] L. Sargeant *et al.*, "Family history of diabetes identifies a group at increased risk for the metabolic consequences of obesity and physical inactivity in epic-norfolk: a population-based study," *Int. J. Obes.*, 2000.
- [33] I. Ikeda *et al.*, "Cross-sectional and longitudinal associations between family history of type 2 diabetes mellitus, hypertension, and dyslipidemia and their prevalence and incidence: Toranomon hospital health management center study (topics24)," *Mayo Clin. Proc.*, 2025.
- [34] R. A. Shaikh *et al.*, "Socioeconomic status, smoking, alcohol use, physical activity, and dietary behavior as determinants of obesity and body mass index in the united states: findings from the national health interview survey," *Int. J. MCH AIDS*, 2015.
- [35] L. Franzini *et al.*, "Neighborhood characteristics favorable to outdoor physical activity: disparities by socioeconomic and racial/ethnic composition," *Health & Place*, 2010.
- [36] V. T. Tong *et al.*, "Age and racial/ethnic disparities in pre-pregnancy smoking among women who delivered live births," *Preventing Chronic Disease*, 2011.
- [37] M. Amiri *et al.*, "Risk of hypertension in women with polycystic ovary syndrome: a systematic review, meta-analysis and meta-regression," *Reprod. Biol. Endocrinol.*, 2020.
- [38] F. Fuchs *et al.*, "Effect of maternal age on the risk of preterm birth: A large cohort study," *PLOS ONE*, 2018.
- [39] A. Girsan *et al.*, "Women's pre-pregnancy underweight as a risk factor for preterm birth: a retrospective study," *BJOG*, 2016.
- [40] C. Boomsma *et al.*, "A meta-analysis of pregnancy outcomes in women with polycystic ovary syndrome," *Human Reproduction Update*, 2006.
- [41] S. D. Sullivan *et al.*, "Hypertension complicating diabetic pregnancies: Pathophysiology, management, and controversies," *J. Clin. Hypertens.*, 2011.
- [42] H. An *et al.*, "Impact of gestational hypertension and pre-eclampsia on preterm birth in china: a large prospective cohort study," *BMJ Open*, 2022.