LLMs for Causal Reasoning in Medicine? A Call for Caution

Saurabh Mathur[1*], **Ranveer Singh**[1*], **Michael Skinner**[1,2], **Predrag Radivojac**[3], **David M. Haas**[4], **Lakshmi Raman**[2] and **Sriraam Natarajan**[1]

[1]The University of Texas at Dallas, Richardson, TX, USA
[2]The University of Texas Southwestern Medical Center, Dallas, TX, USA
[3]Northeastern University, Boston, MA, USA
[4]Indiana University School of Medicine, Indianapolis, IN, USA
{saurabhsanjay.mathur, ranveer.singh, mas140130}@utdallas.edu, predrag@northeastern.edu, dahaas@iu.edu, Lakshmi.Raman@UTSouthwestern.edu, sriraam.natarajan@utdallas.edu

Abstract

Large Language Models (LLMs) are increasingly being used for medical advice by patients and healthcare providers. These models capture knowledge from their training data, which consists of vast medical corpora. However, they lack the ability to use this knowledge to causally reason about the underlying physiological processes. Moreover, they are unable to deal with uncertainty, generating responses that are confidently presented yet factually incorrect. Acting on such factually incorrect medical advice can be dangerous. Mitigating these risks requires rethinking the role of LLMs in medicine. In this work, we present an evaluation scheme for LLMs in three roles: direct clinical decision support, exact medical knowledge base, and approximate medical knowledge base. We evaluate six LLMs on two clinical studies, in obstetrics and pediatric critical care, respectively. Our results indicate that LLMs are much better suited to the approximate knowledge base role. Based on these observations, we request caution when directly employing LLMs in safety-critical domains such as medicine.

1 Introduction

Large language models (LLMs), such as Generative Pretrained Transformer (GPT) and Gemini, have generated significant interest in their potential to assist or even replace aspects of medical practice, with speculation regarding their ability to generate differential diagnoses and treatment plans and especially to reduce administrative burdens [McCoy *et al.*, 2024a]. Indeed, internal medicine residents already perceive a current and future role for LLMs in medicine and use these tools in professional settings, often without formal guidance [Fried *et al.*, 2024]. Moreover, patients are increasingly using LLMs to obtain medical advice [Kohane, 2024]. However, despite the compelling fluency of LLM-generated text, LLMs cannot reason [Zečević *et al.*, 2023; Kambhampati, 2024]. Rather, LLMs are designed to mimic human utterances by identifying linguistic patterns from large corpora; they lack an explicit logical or causal reasoner, a stark contrast to the way clinicians manage patients.

Medical practice is fundamentally rooted in sophisticated cognitive processes, and especially causal reasoning [Kuipers and Kassirer, 1984]. Clinicians develop causal models to understand physiological mechanisms, evaluate hypotheses, construct explanations, and devise physiological interventions. Such causal concepts are very challenging for LLMs; they struggle to perform reasoning tasks not represented in their training data, such as simple math problems involving infrequently used numbers [Yasaman *et al.*, 2022]. In practice, this results in "hallucinations" or "confabulations": coherent and confident yet factually incorrect statements; these are particularly dangerous in a medical context where accuracy is critical. In one report soliciting advice from an LLM to manage a serious infection, the LLM suggested dangerously incorrect management plans contradicting clinical guidelines [Schwartz *et al.*, 2024]. The persuasive nature of LLM outputs can also exploit human automation bias, potentially leading clinicians to over-rely on machine suggestions and make errors. Moreover, inserting factually incorrect text directly into medical records could diminish the quality of information, impede clinical reasoning, and even hinder the development of future AI tools [McCoy *et al.*, 2024a]. Given these substantial risks and limitations in areas critical for causal reasoning, it is dangerous and unethical to rely on current LLMs to diagnose and manage human disease. However, it is likely that less ambitious tasks may be assigned to LLMs to aid in medical practice and clinical decision support.

To this effect, we evaluate the efficacy of pretrained LLMs as approximate sources of causal knowledge, focusing on two clinical studies in obstetrics and pediatric critical care. Specifically, we propose a three-stage evaluation scheme for such systems, consisting of pairwise question answering, full causal graph construction, and the refinability of the constructed causal graph. Our evaluation shows that while LLMs struggle to answer causal questions, their answers can be used as initial hypotheses to construct models more amenable to causal reasoning, such as Causal Bayesian Networks.

[*]Equal contributors

2 Background

2.1 Large Language Models

Large Language Models (LLMs [Minaee *et al.*, 2024]) are a class of generative models that represent the probability distribution over natural language text using neural networks, typically based on the transformer architecture [Vaswani *et al.*, 2017]. They are distinguished from other language models, such as Hidden Markov Models (HMMs [Rabiner and Juang, 1986]), by their unprecedented scale, with some of the largest models boasting hundreds of billions of trainable parameters. This large size allows them to capture intricate statistical patterns from large corpora of natural language text.

LLMs can be categorized into encoder-only, decoder-only, and encoder-decoder models. Of these, decoder-only models have demonstrated impressive capabilities across a wide range of natural language processing tasks. These models process text in one direction, modeling its probability autoregressively, that is, the probability of each word is conditioned on all the words before it. This autoregressive structure allows decoder-only LLMs to be efficiently trained on vast amounts of unlabeled text through simple tasks such as next-word prediction, allowing them to generate highly realistic text. As a result, decoder-only LLMs, including models fine-tuned on medical data, demonstrate impressive performance on medical benchmarks such as question answering, clinical note summarization, patient report generation, and diagnostic reasoning [Luo *et al.*, 2022; Singhal *et al.*, 2025; Schwartz *et al.*, 2024].

While LLMs can generate clinically relevant and accurate text that mimics causal reasoning, they do not perform true causal inference. As Zečević *et al.* (2023) note, LLMs rely on statistical correlations rather than causal understanding, making them prone to blending genuine causal relationships with spurious associations. Moreover, the decoder-only architecture makes them inherently stochastic and prone to cascading errors [McCoy *et al.*, 2024b; Holtzman *et al.*, 2019]. These limitations – an inability to reason, stochasticity, and cascading errors – result in these models generating confident-sounding yet factually incorrect text, especially about topics less represented in training data. This phenomenon is referred to as a hallucination or a confabulation. Since confabulations are the result of inherent limitations of decoder-only LLMs, commonly used mitigation strategies such as Retrieval-Augmented Generation (RAG, Lewis *et al.* [2020]) are inadequate; they may even result in unsafe text generation [An *et al.*, 2025].

2.2 AI-in-the-loop

LLM use in medicine can be analyzed by characterizing the nature of human-AI interaction. Since clinical practice requires the human clinician to be the primary decision-maker [Chin-Yee and Upshur, 2018], it is an example of an AI-in-the-loop domain [Natarajan *et al.*, 2025]. Here, the AI system's role is to support the clinician by providing them with accurate and actionable information. The efficacy of such a system depends on its ability to improve the clinician's decision-making, such as by alleviating their cognitive load by automating mechanical aspects of clinical reasoning.

However, LLM use in medicine can potentially deviate from the AI-in-the-loop paradigm. Fig. 1 depicts this deviation. Clinicians could treat LLMs as an expert system, asking them questions that typically require causal inference. Since LLMs are unable to reason, stochastic, and prone to generating erroneous output, the human clinician would need to validate the LLM-generated answers, increasing their burden [Karabacak and Margetis, 2023]. Moreover, since the erroneous output is often confidently phrased, there is a distinct possibility that clinicians under pressure might miss one or more errors, which might carry over in their ultimate decision. Mitigating these risks requires a fundamental rethinking of the role LLMs might play in clinical decision support.

2.3 Causal Bayesian Networks

LLMs can be contrasted with another class of generative models called Causal Bayesian Networks (CBNs, Pearl [2009]); unlike LLMs, CBNs are interpretable, amenable to causal reasoning, and naturally deal with uncertainty, satisfying the desiderata for AI-in-the-loop in medicine. CBNs are closely related to causal diagrams, which clinicians have used for causal reasoning [Kuipers and Kassirer, 1984].

CBNs are a subclass of Bayesian Networks (BNs). BNs represent the joint probability distribution over a set of variables by factorizing it over a directed acyclic graph (DAG). This DAG consists of nodes corresponding to each variable; each directed edge between two variables denotes direct influence. If the edges also denote direct causal relationships, then the BN is considered a CBN. In a CBN, each edge $X \rightarrow Y$ means that X is a cause of Y. These causal edges can be interpreted interventionally: intervening on the component corresponding to X should change the distribution over Y. Such targeted interventions might not be reasonable in some cases, such as when modeling the effect of the Family history of a medical condition; in such cases, the edges can be interpreted historically or etiologically [Glymour and Glymour, 2014].

CBNs can be constructed by eliciting them from domain experts or clinical guidelines such as the Quick Medical Reference (QMR Shwe and others [1991]). This approach falls short when modeling medical conditions that are less well-understood, such as rare diseases and conditions involving complex causal relationships. As a result, considerable research has been performed to devise ways to generate causal graphs from observational data [Guo *et al.*, 2020].

Data-driven causal discovery methods typically use large amounts of data to exclude non-causal edges and rely on several assumptions to decide the causal direction of the remaining edges; examples of such methods include Peter-Clark (PC, Spirtes *et al.* [2000]), Greedy Equivalence Search (GES, Koller and Friedman [2009]), and Fast Causal Inference (FCI, Spirtes *et al.* [2000]). The assumptions used by these methods include the causal Markov condition, which states that each variable is independent of its non-effects (non-descendants) given its direct causes; faithfulness, which states that any conditional independencies in the data arise from the structure of the causal graph itself; and causal sufficiency, which requires all common causes of observed variables to be included in the dataset.

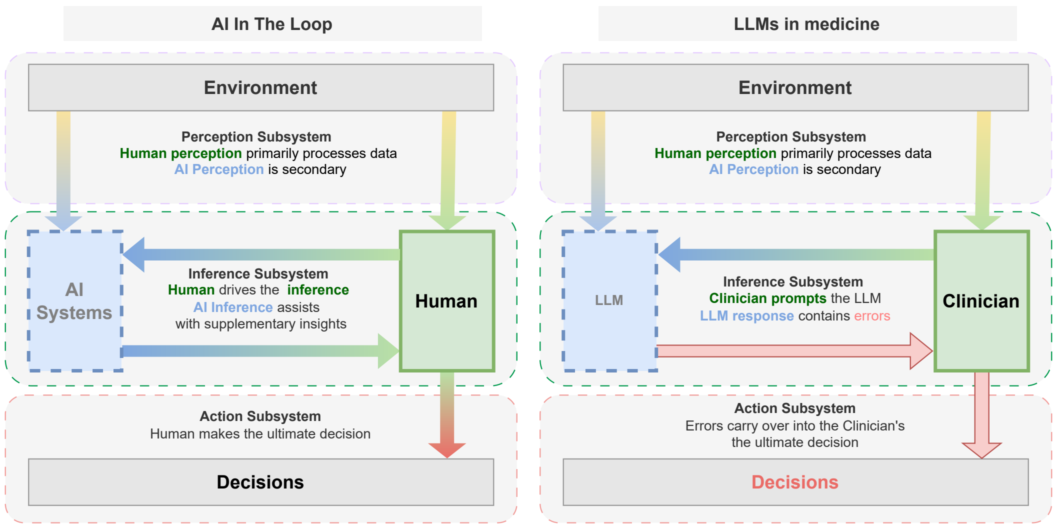

Figure 1: In AI-in-the-loop domains (left), humans must make the ultimate decisions, while AI systems assist with perception, inference, and action. The fluency of LLM-generated text blurs the lines between AI assistance and human decision-making. As a result, using LLMs as AI-in-the-loop systems in medicine (right) risks introducing new errors to clinical practice. In response to a difficult query, the LLM might generate confident-sounding text with errors that the human clinician might fail to catch, carrying them over in their final decision.

Data-driven causal discovery in medicine is challenging. First, the complex dynamics underlying medical domains make it difficult to make assumptions like causal sufficiency without expert knowledge. Second, human physiology is naturally cyclic, while causal graphs represent acyclic relationships [Claassen *et al.*, 2021]. This requires the causal modeling domain and its variables to be carefully designed. Finally, expensive data annotation, the temporal nature of the data, missingness of variables due to data collection issues further complicate data-driven causal discovery in medicine [Sanchez *et al.*, 2022; Kitson *et al.*, 2023; Zanga *et al.*, 2022]. CBN construction in medical domains requires a hybrid approach, combining expert domain knowledge with empirical patterns.

2.4 Theory Refinement

Theory refinement is one such hybrid method that combines expert knowledge with data-driven learning. In it, an expert-specified model is refined to better fit empirical evidence [Mooney and Shavlik, 2021]. This method has been used to improve the structure of BNs derived from incomplete or imperfect domain knowledge [Buntine, 1991], such as LLM output [Mathur *et al.*, 2024; Mathur *et al.*, 2025]. Refining a BN involves adding, removing, or reversing edges from it to maximize a score measuring the empirical validity of the relationships between variables. Commonly used scores include Bayesian-Dirichlet (BD, Heckerman *et al.* [1995]) and the Minimal Description Length (MDL, Lam and Bacchus [1993]) scores. The MDL score has been used in causal discovery since it approximates the Kolmogorov complexity of the causal graph [Janzing and Schölkopf, 2010; Mian *et al.*, 2023]. It consist of two components — the log-likelihood of the data under the CBN ($\mathcal{L}(\mathcal{M}, \mathcal{D})$) and the description length cost of the CBN ($Cost(\mathcal{M})$), Score($\mathcal{M}$) $= \mathcal{L}(\mathcal{M}, \mathcal{D}) - Cost(\mathcal{M})$.

3 Assessment of LLMs for Causal Reasoning

LLMs excel at capturing statistical patterns from large amounts of textual data, allowing them to synthesize highly coherent text and achieve impressive performance on medical benchmarks. This has generated significant interest in using them for clinical decision support, as AI-in-the-loop. However, these models are stochastic and lack a reasoner, making their responses a blend of accurate and inaccurate information. We aim to empirically evaluate the performance of models on real-world medical domains.

3.1 Data sets

We consider two medical domains: nuMoM2b and PELICAN. nuMoM2b is an obstetrics domain, based on a study that aims to understand Adverse Pregnancy Outcomes (APOs) in nulliparous subjects (first-time mothers). APOs are common, affecting 15% of U.S. pregnancies, and the study covers a time scale of 8 to 9 months. In contrast, PELICAN is based on a pediatric critical care study that aims to understand neurological injury in pediatric subjects supported by Extracorporeal Membrane Oxygenation (ECMO). This condition is extremely rare, and the study deals with a much shorter period, encompassing the duration during which the patient was on life support (less than one month). The dataset size for PELICAN is considerably smaller, with only 71 subjects, and existing research on neurological injury in pediatric patients on ECMO is less abundant, with 34,000 results on Google Scholar, 80 times fewer results than APOs. Table 1 summarizes the differences between the two domains, and tables 3 and 2 summarize the specific variables considered for our evaluation. These variables were selected by our domain experts.

	nuMoM2b	PELICAN
Subfield	Obstetrics	Pediatric Critical Care
Time scale	8 to 9 months	$<$ 1 month
Condition(s) considered	Adverse pregnancy outcomes	Neurological injury on life support
Rarity of condition(s)	Common (15% of US pregnancies)	Extremely rare (20% out of $<$ 2500 cases a year in the US)
Existing research	2.7M results on google scholar	34k results on google scholar
Subject inclusion criteria	First-time mothers (Nulliparous) without pregestational diabetes	Pediatric patients supported by ECMO, but not having congenital heart disease
Average age of subjects	27.79 years	4.32 years
Risk factors	Demographics, existing conditions, family history, and lifestyle factors recorded at start of pregnancy	Abnormal events identified from high-frequency physiological measurements and laboratory test results up to 24 hours on ECMO
Data set size	3,856	71

Table 1: Comparison of the two medical domains and the corresponding subsets considered in this work

Variable	Value	% subjects
Age	≤ 21	15.9%
	21-35	77.0%
	> 35	7.0%
BMI	≤ 18	1.3%
	18-25	54.1%
	> 25	44.6%
Race	Non-Hispanic Asian	4.2%
	Non-Hispanic Black	11.2%
	Non-Hispanic White	67.8%
	Hispanic	12.3%
	Others	4.43%
DiabHist	TRUE	20.7%
HTNHist	TRUE	45.4%
HiBP	TRUE	2.7%
PCOS	TRUE	4.8%
METS	TRUE	66.1%
Smoking	TRUE	14.9%
PReEc	TRUE	5.9%
NewHTN	TRUE	17.7%
PTB	TRUE	7.7%
GDM	TRUE	3.8%
Total		3,856

Table 2: **Risk Factors and Outcomes for nuMoM2b.** We consider four adverse outcomes: Preeclampsia (PReEc), New Hypertension (NewHTN), Gestational Diabetes Mellitus (GDM), and Preterm Birth (PTB). For these adverse outcomes, we consider nine risk factors: Age, Body Mass Index (BMI), Race, Family History of Diabetes and Hypertension (DiabHist and HTNHist, respectively), Hypertension (HiBP), Polycystic Ovary Syndrome (PCOS), physical activity measured in Metabolic Equivalents of Time (METs), and Smoking in the three months before start of pregnancy.

Variable	Value	% subjects
HighVIS	TRUE	21.1%
Hypotension	TRUE	23.9%
Hypertension	TRUE	4.2%
LowPlatelet	TRUE	32.4%
HighLactate	TRUE	59.2%
LowpH	TRUE	9.86%
RelativepCO2	TRUE	29.6%
NeurologicalInjury	TRUE	23.9%
Total		71

Table 3: **Risk Factors and Outcomes for PELICAN**. We consider the adverse outcome of Neurological Injury. We consider seven of its risk factors: High Vasoactive-inotropic score (HighVIS), Hypotension, Hypertension, Low Platelets, High Lactate, Low pH, as well as the high relative change of pCO2 24 hours post-canulation, as compared to pre-canulation levels.

3.2 Evaluation scheme

To empirically assess each LLM's performance on these real-world medical domains, we use a three-stage evaluation scheme. First, to assess the LLM's ability to answer **direct causal questions**, we prompt it with queries about every pair of variables. We use their answers to construct a causal graph and evaluate this graph. For the second stage, we evaluate the **LLM as a knowledge source** and prompt it to construct a full causal graph from the given list of variables. Finally, the third stage evaluates the **LLM as an approximate knowledge source** , focusing on the refinability of the LLM-constructed causal graph. Here, we combine the LLM-generated graph with indirect expert knowledge like anticausal relations based on temporal order, and refine it using empirical data to further eliminate incorrect edges. We evaluate this refined graph.

Each of these graphs is compared against graphs constructed by our domain experts. Since both domains are being actively researched, these expert graphs do not fully capture all the causal relationships, but they do capture known causal relationships, to the best of our experts' knowledge.

3.3 Metrics

We evaluate the structure of causal Bayesian networks (CBNs) by comparing them against expert-provided causal graphs. To quantify the difference, we use three metrics: Structural Hamming Distance (SHD [Acid and De Campos,

```
Extracorporeal Membrane
Oxygenation, or ECMO for short, is
an advanced therapy that is
sometimes used to work the heart
and lungs when a patient's organs
are too sick or weak to work on
their own.  It is effectively a
modified heart-lung bypass
machine-a machine that takes over
heart and lung function (meaning it
adds oxygen to and removes carbon
dioxide from a patient's blood
supply).
For such patients, is there a
causal edge from LowMAP to HighMAP?
Here, LowMAP is hypotension in the
first 24 hours of the ECMO run, and
HighMAP is hypertension in the
first 24 hours of the ECMO run.
Provide the answer as a single
word, Yes or No (with No also for
cases when the answer is unknown).
```

```
## Setup
You are a knowledge engineer working on a study
on mitigating the risk of Adverse Pregnancy
Outcomes (APOs).  Think carefully and
logically, explaining the reasons for your
answer.
## Available Information
Your team has collected variables representing
clinical and demographic information.  The
study data is collected on Nulliparous women
and consists of variables representing clinical
and demographic features that might influence
the risk of Adverse Pregnancy outcomes.  The
variables being considered are as follows:
{Variable_Descriptions}
## Task
Make a list of direct causal relations between
these variables and explain the rationale
behind each decision.  Please provide the
answers in the format:  Variable 1 -> Variable
2, followed by an explanation on a different
line.
```

Figure 2: Prompts used for pairwise (left) and full causal graph elicitation (right) for PELICAN and nuMoM2b, respectively

2003]), Structural Intervention Distance (SID [Peters and Bühlmann, 2015]), and the number of spurious edges (SE). SHD quantifies the number of edge additions, deletions, or reversals needed to convert the learned graph into the true graph. SID captures the number of incorrect inferences a learned graph makes about intervention effects compared to the true causal model. Finally, SE refers to the edges present in the learned graph but not in the true graph.

4 Results of Empirical Evaluation

We now present the results of our empirical investigation[1] and try to answer the following questions

1. **How do LLMs perform as causal question answering systems?**
2. **How do LLMs perform as exact knowledge bases?**
3. **How do LLMs perform as approximate knowledge bases?**

We evaluate our results using 10 bootstrap samples for each dataset. To construct each LLM's representative causal graph, the model was prompted five times with the same full prompt for the full-graph generation. For the question-based graph construction, each question was similarly posed five times. These responses were aggregated to construct a DAG, adding edges in decreasing order of frequency across the five runs, excluding any edge that would introduce a cycle. Ties in edge frequency were resolved lexicographically based on the name of the source node.

[1]Please refer to the supplementary material for additional details of the experimental setup, including data preprocessing, LLM prompts, and responses: https://github.com/s-ranveer/LLM-Causal-Medicine-Eval

Results are presented in Table 5 for six LLMs: Claude [Anthropic, 2023], DeepSeek [Liu *et al.*, 2024], Gemini [Team *et al.*, 2023], GPT-4o [OpenAI, 2025], LLaMA [Touvron *et al.*, 2023], and the LLM fine-tuned on medical text, OpenBioLLM [Liu *et al.*, 2025]. For the data-only baselines, we consider the constraint-based Peter-Clark (PC) algorithm that starts with a fully connected undirected graph and uses statistical independence tests to remove or orient edges. Additionally, we consider the score-based Greedy Search and Score (GSS) algorithm, which is based on Greedy Equivalence Search (GES), where we evaluate graph structures by optimizing a score function such as the Bayesian Information Criterion (BIC). Finally, we consider Fast Causal Inference (FCI), which is designed to handle latent confounders and learn causal features that remain consistent across all graphs in an equivalence class.

4.1 LLMs for Causal Question Answering

To evaluate whether LLMs can be used for causal reasoning in obstetrics and pediatric critical care, we posed questions about the causal relationships between pairs of variables to the LLMs using the prompt illustrated in Figure 2 (left). The LLM responses were then compared to an expert-constructed causal graph. Table 5 (middle and bottom) presents the experimental results.

Both domains present unique challenges to the LLMs. For pediatric critical care, the limited availability of relevant literature on ECMO likely restricts the LLM's ability to identify accurate causal relationships from its training data. Conversely, the obstetrics domain has a lot more literature, but it involves a significantly larger number of variables. As a result, asking pairwise causal questions without sufficient contextual information results in spurious associations, driven by

Domain	LLM	Deleted/Total	
		Pairwise	Full
PELICAN	Claude	0/4	2/13
	Deepseek	1/13	3/17
	Gemini	13/35	2/17
	GPT 4o	1/10	1/11
	LLaMA	20/46	7/25
	OpenBioLLM	17/43	1/9
nuMoM2b	Claude	1/18	0/32
	Deepseek	0/27	0/31
	Gemini	0/40	0/34
	GPT 4o	1/37	0/25
	LLaMA	32/95	1/32
	OpenBioLLM	39/99	2/43

Table 4: The number of edges deleted to eliminate cycles during causal graph construction, for both prompt types and across both domains.

hidden confounders. This issue is evident in the pairwise results for both domains, where the performance is close to purely data-driven baselines. The best-performing models in these scenarios tend to be those that responded more conservatively, affirming fewer causal relationships like Claude and Deepseek, as seen in table 4, resulting in lower SHD, SID, and SE values.

Additionally, the prompt responses often exhibit inconsistencies in determining the direction of causal relationships between variable pairs. When asked using the pairwise prompt format, LLMs frequently respond affirmatively to both directions—i.e., $A \rightarrow B$ and $B \rightarrow A$—thereby introducing cycles. This can be seen in Table 4, which presents the number of edges deleted to enforce acyclicity for each case. Therefore, LLMs by themselves perform poorly as casual question answering systems.

4.2 LLMs as Exact Knowledge Bases

To evaluate the potential of LLMs as medical knowledge bases, we provided each model with a prompt for one-shot full causal graph construction, including the domain description, variable definitions, and the overall task description. As in the pairwise evaluation, the generated graphs were compared against expert-established causal structures. Table 5 (middle and bottom) presents the results of full causal graph construction in both domains.

While the domain-specific challenges discussed earlier are not entirely resolved by providing the full set of available variables, they are significantly mitigated. As a result, LLM performance improves notably when using the full prompt compared to pairwise prompting. As in pairwise prompting, conservative models, such as Claude and DeepSeek, produced fewer edges, performing better than the other models.

Full prompt-based causal graphs outperform data-driven causal discovery methods. However, despite these improvements, the number of spurious edges, SHD, and SID remains too high to fully trust the LLM-generated graphs. Therefore, while LLMs show promise, they cannot yet be relied upon as standalone exact medical knowledge bases.

4.3 LLMs as Approximate Knowledge Bases

To evaluate the use of LLMs as approximate knowledge bases, we refine their outputs using data and indirect domain knowledge about temporally impossible edges. The refinement procedure deletes edges to maximize the MDL score. Table 5 (middle and bottom) shows the results after refining the LLM-generated graphs.

For the pediatric critical care domain, we see a reduction in the SHD and the number of spurious edges (SE) across most LLMs in both pairwise and full prompting, with a reduction in SID for some of the LLMs. However, in the obstetrics domain, performance improves only for the pairwise prompt. For the full prompt, only OpenBioLLM shows significant improvement, while the others have similar or slightly worse performance. This indicates the lower refinability of LLM-generated causal graphs in the obstetrics domain.

Overall, refinement appears to be more effective for the PELICAN domain than for nuMoM2b. This is likely due to the relatively limited literature available on pediatric critical care, which limits the LLM's exposure during training. As a result, the model is more prone to generate non-causal edges that are non-associational, and hence easier to remove through the refinement process. In contrast, obstetrics is a well-studied domain with a broad body of research covering diverse populations. Moreover, obstetrics-related discussions are more prevalent in public discourse, often drawing from a mix of high- and low-quality sources. As a result, LLM outputs in this domain may include specific causal claims that are either not credible or not applicable to the nuMoM2b study population. This can lead to suboptimal refinement, including the unintended removal of valid edges from the graph.

Despite these limitations, the graphs constructed from LLM-responses are more accurate than those discovered from the limited data using algorithms like GSS, PC, and FCI, as seen in table 5 (Top). The difference becomes more pronounced after refinement, especially for LLMs that output a lot of causal edges like OpenBioLLM and LLaMA. Therefore, LLMs have utility as approximate knowledge sources.

5 Conclusion

We considered LLM use in medical practice. These models capture intricate statistical patterns from vast medical corpora to generate fluent text, achieving high performance on medical benchmarks. Indeed, there has been significant interest in their potential to assist or even replace aspects of medical practice. However, their lack of an explicit causal reasoner, along with their stochasticity, raises concerns about their suitability as clinical decision support systems. We proposed an evaluation scheme to evaluate LLMs in three different roles in the clinical decision support pipeline. We evaluated six LLMs on two medical domains. Our results indicate that while LLMs do capture medical domain knowledge from their training data, they fail to accurately answer causal questions. LLM-use requires caution, especially in high-stakes domains like medicine, but these models might be used as approximate knowledge sources to construct models more amenable to causal reasoning, like Causal Bayesian Networks. There are a number of directions for future work.

Baseline	PELICAN			nuMoM2b		
	SHD	SID	SE	SHD	SID	SE
Greedy Search and Score (GSS)	9.5 ± 1.6	19.2 ± 4.6	4.2 ± 1.5	33 ± 1.5	90 ± 5.5	10.8 ± 1.1
Peter and Clarke (PC)	8.5 ± 1.1	18.7 ± 4.0	1.2 ± 1.0	33.7 ± 1.7	91.9 ± 7.5	7.5 ± 2.4
Fast Causal Inference (FCI)	8.0 ± 0.5	14.9 ± 0.3	0.1 ± 0.3	31.8 ± 1.2	80.6 ± 4.1	2.3 ± 1.7

	LLM	LLM output			Subtractive refinement		
		SHD	SID	SE	SHD	SID	SE
Pairwise	Claude	6	6	1	6.3 ± 0.5	11.9 ± 1.2	0
	DeepSeek	9	7	7	8 ± 1.2	13.4 ± 2.2	1.8 ± 1.1
	Gemini	21	8	18	14.3 ± 1.6	11.6 ± 1.6	9 ± 1.5
	GPT 4o	9	14	6	6.6 ± 0.9	12.8 ± 1.9	0.4 ± 0.5
	LLaMA	24	17	23	14.3 ± 1.4	16.7 ± 3.9	8.7 ± 1.2
	OpenBioLLM	23	11	22	14.8 ± 3	14.6 ± 2.5	8.5 ± 2.5
Full	Claude	4	5	4	4.5 ± 0.7	5 ± 1.2	4.8 ± 1.1
	DeepSeek	6	0	6	4.8 ± 0.9	4.2 ± 1.3	3.2 ± 0.4
	Gemini	9	10	9	5.9 ± 1.0	6.5 ± 2.4	3.5 ± 0.7
	GPT 4o	8	15	6	7 ± 1.0	12 ± 3.7	1.6 ± 0.9
	LLaMA	14	12	13	8.3 ± 1.0	7.1 ± 1.7	5.1 ± 0.9
	OpenBioLLM	9	19	5	7.8 ± 1.2	14.3 ± 1.5	0.9 ± 0.9

	LLM	LLM output			Subtractive refinement		
		SHD	SID	SE	SHD	SID	SE
Pairwise	Claude	23	63	6	27.5 ± 0.8	66.9 ± 0.7	4.8 ± 0.6
	DeepSeek	23	56	10	22.1 ± 0.7	56.3 ± 1.2	7.5 ± 0.5
	Gemini	32	49	21	30 ± 0.5	49.7 ± 1.1	17.6 ± 4.8
	GPT 4o	25	45	15	23.8 ± 0.4	41.6 ± 1.8	13.7 ± 0.5
	LLaMA	45	52	41	24.6 ± 0.9	37.9 ± 3.3	15.8 ± 1.3
	OpenBioLLM	43	35	38	31.8 ± 1.0	26.7 ± 0.6	23.5 ± 0.5
Full	Claude	17	44	9	18.5 ± 0.8	49.5 ± 0.8	8.2 ± 0.6
	DeepSeek	21	50	10	22.7 ± 1.3	53.1 ± 1.1	9.3 ± 1.0
	Gemini	16	38	9	16.1 ± 0.9	39.2 ± 1.7	8.7 ± 0.6
	GPT 4o	20	53	5	21.9 ± 0.8	58.5 ± 3.4	4.1 ± 0.3
	LLaMA	26	53	11	26.7 ± 1.5	55.6 ± 1.5	9.6 ± 0.5
	OpenBioLLM	32	54	22	26.2 ± 1.0	45 ± 4.4	13.5 ± 0.5

Table 5: Evaluation results comparing graphs constructed by each method to corresponding expert graphs; the difference is quantified in terms of Structural Hamming Distance (SHD), Structural Interventional Distance (SID), and the number of spurious edges (SE). The tables show results for the three data-driven baselines on both data sets (top), and the LLM-generated graphs on the PELICAN (middle) and nuMoM2b (bottom) domains.

First, this evaluation can be expanded to more medical domains. Second, imposing validity constraints on the LLM responses can improve the quality of LLM-generated graphs. Finally, multiple LLMs might be combined to create a more reliable ensemble of approximate knowledge sources.

6 Acknowledgements

We acknowledge support from NIH award R01NS133142.

References

[Acid and De Campos, 2003] S. Acid and L. M. De Campos. Searching for Bayesian network structures in the space of restricted acyclic partially directed graphs. *JAIR*, 18:445–490, May 2003.

[An *et al.*, 2025] Bang An, Shiyue Zhang, and Mark Dredze. Rag llms are not safer: A safety analysis of retrieval-augmented generation for large language models, 2025.

[Anthropic, 2023] Anthropic. Claude: An ai assistant. https://www.anthropic.com/claude, 2023. Accessed: 2025-05-26.

[Buntine, 1991] Wray Buntine. Theory refinement on bayesian networks. In *UAI*, 1991.

[Chin-Yee and Upshur, 2018] Benjamin Chin-Yee and Ross Upshur. Clinical judgement in the era of big data and predictive analytics. *Journal of Evaluation in Clinical Practice*, 24(3):638–645, 2018.

[Claassen *et al.*, 2021] Jurgen AHR Claassen, Dick HJ Thijssen, Ronney B Panerai, and Frank M Faraci. Regulation of cerebral blood flow in humans: physiology and clinical implications of autoregulation. *Physiological reviews*, 2021.

[Fried *et al.*, 2024] Aaron J Fried, Spencer D Dorn, William J Leland, Emily Mullen, Donna M Williams, Aimee K Zaas, Jack MacGuire, and Debra L Bynum. Large language models in internal medicine residency: current use and attitudes among internal medicine residents. *Discover Artificial Intelligence*, 4(1):70, 2024.

[Glymour and Glymour, 2014] Clark Glymour and Madelyn R Glymour. Commentary: race and sex are causes. *Epidemiology*, 25(4):488–490, 2014.

[Guo *et al.*, 2020] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4), 2020.

[Heckerman *et al.*, 1995] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20, 1995.

[Holtzman *et al.*, 2019] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[Janzing and Schölkopf, 2010] Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

[Kambhampati, 2024] Subbarao Kambhampati. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1):15–18, 2024.

[Karabacak and Margetis, 2023] Mert Karabacak and Konstantinos Margetis. Embracing large language models for medical applications: opportunities and challenges. *Cureus*, 15(5), 2023.

[Kitson *et al.*, 2023] Neville Kenneth Kitson, Anthony C Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. A survey of bayesian network structure learning. *Artificial Intelligence Review*, 56(8), 2023.

[Kohane, 2024] Isaac S Kohane. Compared with what? measuring ai against the health care we have. *New England Journal of Medicine*, 391(17):1564–1566, 2024.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models*. MIT press, 2009.

[Kuipers and Kassirer, 1984] Benjamin Kuipers and Jerome P Kassirer. Causal reasoning in medicine: analysis of a protocol. *Cognitive Science*, 8(4):363–385, 1984.

[Lam and Bacchus, 1993] Wai Lam and Fahiem Bacchus. Using causal information and local measures to learn bayesian networks. In *UAI*, 1993.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

[Liu *et al.*, 2024] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[Liu *et al.*, 2025] Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Yining Hua, Peilin Zhou, et al. Application of large language models in medicine. *Nature Reviews Bioengineering*, pages 1–20, 2025.

[Luo *et al.*, 2022] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6), 2022.

[Mathur *et al.*, 2024] Saurabh Mathur, Veerendra P Gadekar, Rashika Ramola, Peixin Wang, Ramachandran Thiruvengadam, David M Haas, Shinjini Bhatnagar, Nitya Wadhwa, Garbhini Study Group, Predrag Radivojac, Himanshu Sinha, Kristian Kersting, and Sriraam Natarajan. Modeling Multiple Adverse Pregnancy Outcomes: Learning from Diverse Data Sources. In *AIME*, 2024.

[Mathur *et al.*, 2025] Saurabh Mathur, Ranveer Singh, Michael Skinner, Ethan Sanford, Neel Shah, Phillip Reeder, Lakshmi Raman, and Sriraam Natarajan. Llm-guided causal bayesian network construction for pediatric patients on ecmo. In *International Conference on Artificial Intelligence in Medicine*, pages 255–260. Springer, 2025.

[McCoy *et al.*, 2024a] Liam G McCoy, Arjun K Manrai, and Adam Rodman. Large language models and the degradation of the medical record. *The New England journal of medicine*, 391(17):1561–1564, 2024.

[McCoy *et al.*, 2024b] R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024.

[Mian *et al.*, 2023] Osman Mian, Michael Kamp, and Jilles Vreeken. Information-theoretic causal discovery and intervention detection over multiple environments. In *AAAI*, volume 37, 2023.

[Minaee *et al.*, 2024] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[Mooney and Shavlik, 2021] Raymond J Mooney and Jude W Shavlik. A Recap of Early Work on Theory and Knowledge Refinement. In *AAAI Spring Symposium*, 2021.

[Natarajan *et al.*, 2025] Sriraam Natarajan, Saurabh Mathur, Sahil Sidheekh, Wolfgang Stammer, and Kristian Kersting. Human-in-the-loop or ai-in-the-loop? automate or collaborate? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28594–28600, 2025.

[OpenAI, 2025] OpenAI. Chatgpt: GPT-4o language model. https://openai.com/chatgpt, 2025.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[Peters and Bühlmann, 2015] Jonas Peters and Peter Bühlmann. Structural Intervention Distance for Evaluating Causal Graphs. *Neural Computation*, 27(3):771–799, 2015.

[Rabiner and Juang, 1986] Lawrence Rabiner and Biinghwang Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[Sanchez *et al.*, 2022] Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

[Schwartz *et al.*, 2024] Ilan S Schwartz, Katherine E Link, Roxana Daneshjou, and Nicolás Cortés-Penfield. Black box warning: large language models and the future of infectious diseases consultation. *Clinical infectious diseases*, 78(4):860–866, 2024.

[Shwe and others, 1991] Michael A Shwe et al. Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base. *Methods of information in Medicine*, 30(04), 1991.

[Singhal *et al.*, 2025] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[Team *et al.*, 2023] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017.

[Yasaman *et al.*, 2022] Razeghi Yasaman, Robert Logan IV, Gardner Matt, and S Sameer. Impact of pretraining term frequencies on few-shot numerical reasoning. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, 2022.

[Zanga *et al.*, 2022] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.

[Zečević *et al.*, 2023] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023.