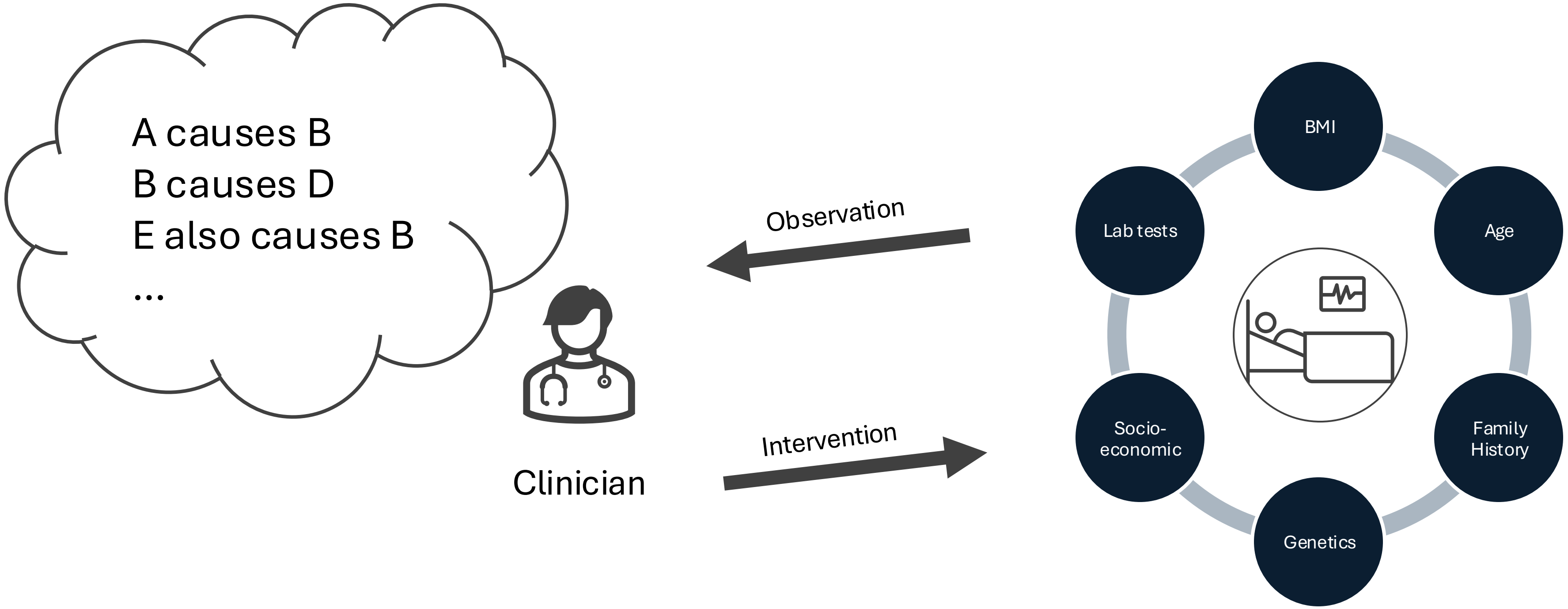
LLMs for Causal Reasoning in Medicine? A Call for Caution

Saurabh Mathur*, Ranveer Singh*, Michael Skinner, Predrag Radivojac, David M. Haas, Lakshmi Raman,
Sriraam Natarajan

*Equal contributors

IJCAI 2025 Workshop on User-Aligned Assessment of Adaptive AI Systems

Clinical practice involves causal reasoning

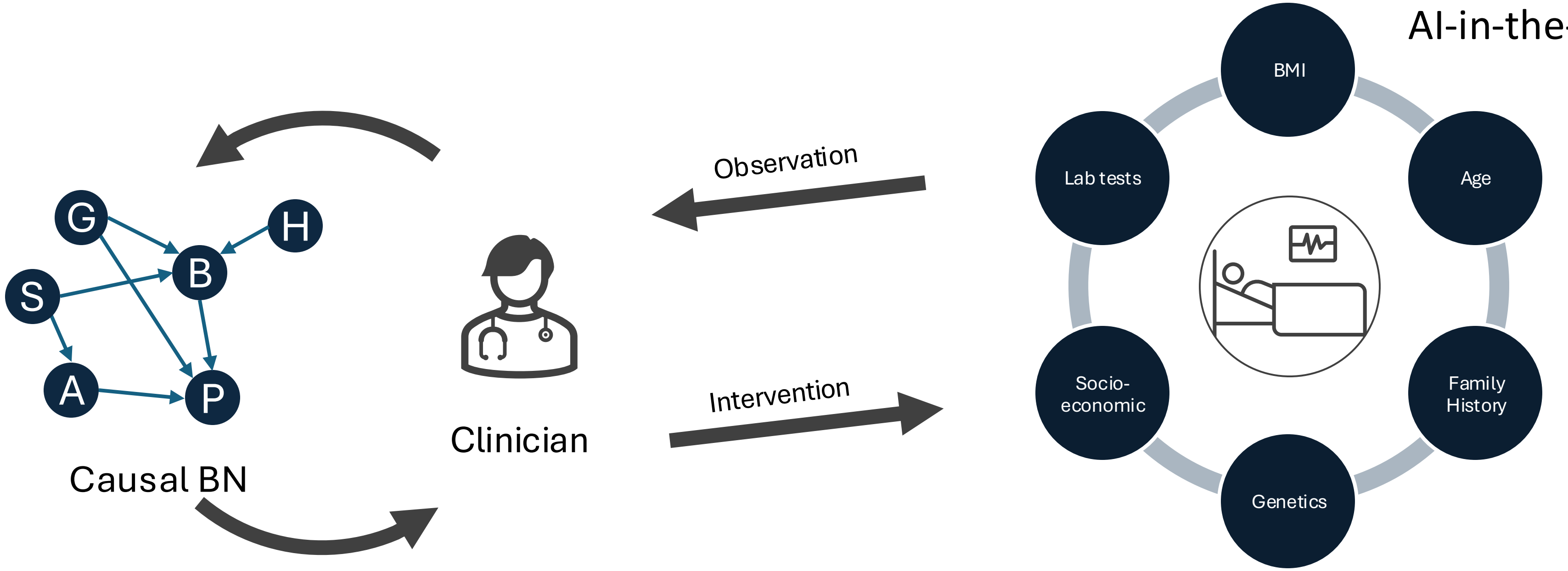


… which can be cognitively taxing.

Kuipers, Benjamin, and Jerome P. Kassirer. "Causal reasoning in medicine: analysis of a protocol." Cognitive Science 8.4 (1984)

Halford, Graeme S. et al. "How many variables can humans process?." Psych. Sci. (2005)

Clinical practice involves causal reasoning



AI-in-the-loop can alleviate cognitive load

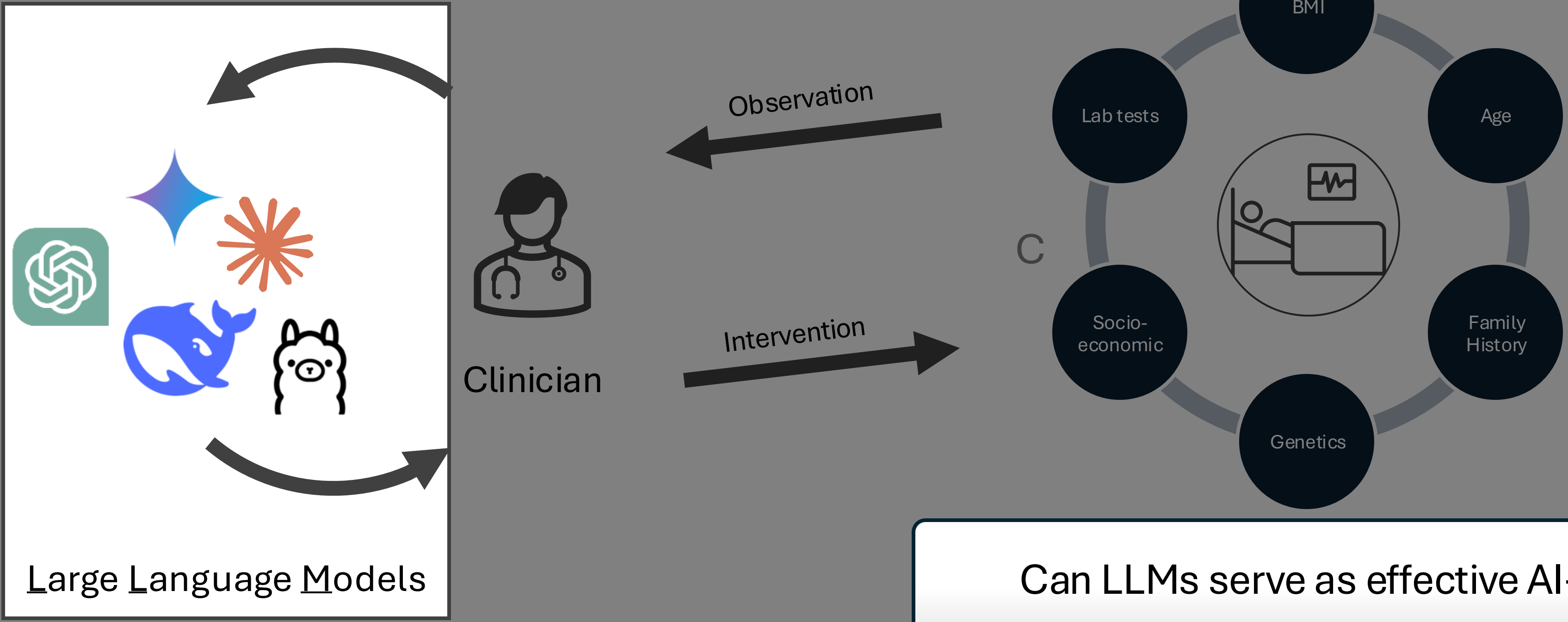
- Making large volumes of data comprehensible
- Keeping up with latest clinical research and guidelines
- Automate mechanical aspects of clinical reasoning

… which can be cognitively taxing.

Natarajan et al., Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?. AAAI (2025)

Clinical practice involves causal reasoning

"26% [internal medicine residents] reported current use."



AI-in-the-loop can alleviate cognitive load

- Making large volumes of data comprehensible
- Keeping up with latest clinical research and guidelines
- Automate mechanical aspects of clinical reasoning

Can LLMs serve as effective AI-in-the-loop?

Natarajan et al., Human-in-the-loop or AI-in-the-loop? Automate or Colla

Fried, Aaron J., et al. "Large language models in internal medicine residency: current use and attitudes among internal medicine residents." *Discover Artificial Intelligence* 4.1 (2024): 70.

Two medical domains

Study	**nuMoM2b**	**PELICAN**
Sub-field	Obstetrics	Pediatric critical care
Subjects	First-time mothers	Severely ill children, supported by ECMO
Condition	Adverse pregnancy outcomes (e.g., preterm birth)	Neurological injury
Rarity of condition	15% of US pregnancies	20% of US ECMO cases, which are < 2.5/year
Time scale	8–9 months	<1 month
Available research	~2.7M publications	~34k publications

Direct causal Q/A: “Is X a cause of Y?”

Question:
Is Hypertension at the start of the pregnancy a cause of New Hypertension during the pregnancy?

Clinician

Response:
Yes, the presence of hypertension at the beginning of pregnancy can increase the risk of developing new hypertension during pregnancy.

OpenBioLLM

Fluent but wrong response

Cascading Errors

Initial mistakes can get amplified through subsequent token generation

Direct causal Q/A: “Is X a cause of Y?”

Clinician

Question:
Is Hypotension a cause of High Lactate for Patients on ECMO?

Claude

Response:
Yes

Claude

Response:
No

Claude

Response:
Yes

Claude

Response:
No

Claude

Response:
Yes

Stochasticity

Inconsistency across outputs for the same prompt.

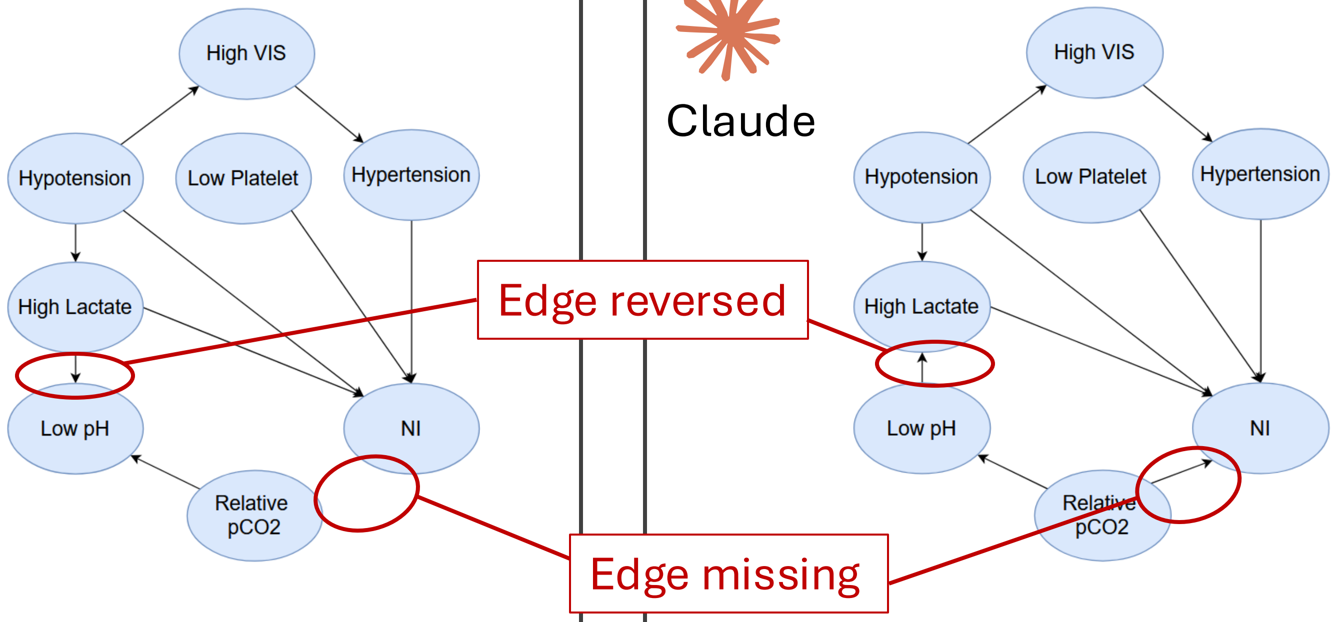
Full causal graph: “Make a graph with these variables”

Clinician

Make a causal graph for ECMO, using these variables [...]

Claude

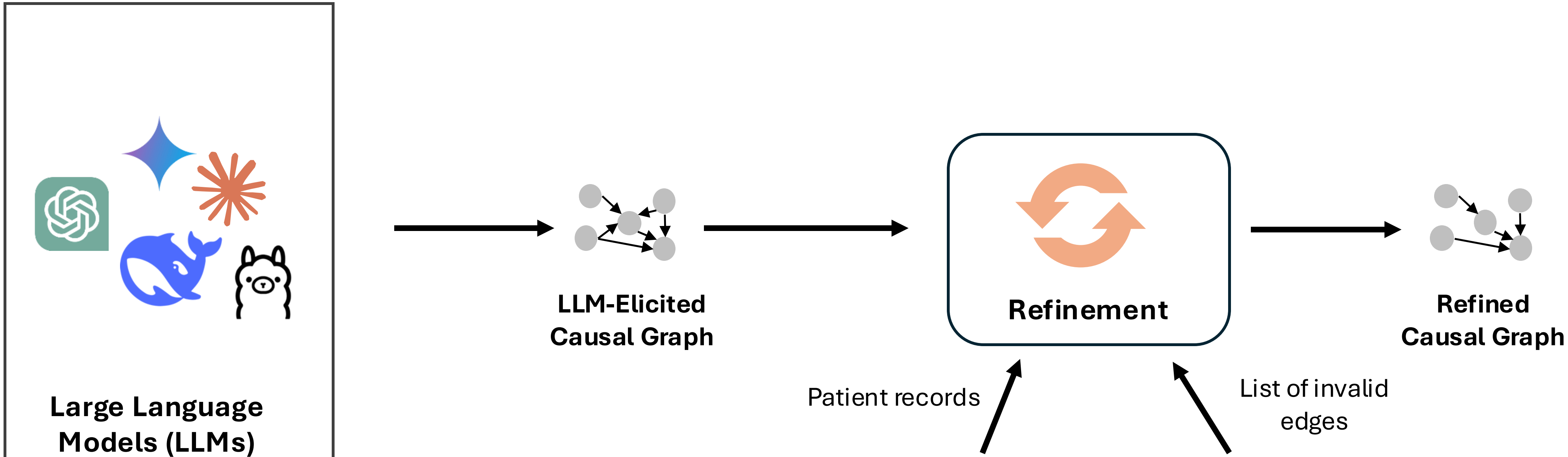
Claude

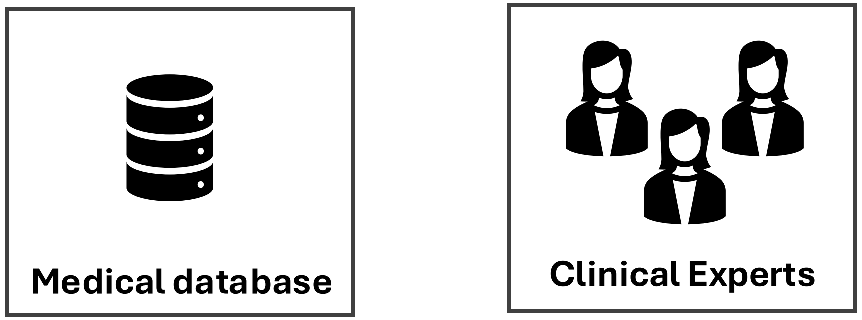


Stochasticity

Inconsistency across outputs for the same prompt.

Theory refinement for LLM-generated graphs





Mathur et al. "Modeling multiple adverse pregnancy outcomes: Learning from diverse data sources." AIME (2024).

Mathur et al. "LLM-Guided Causal Bayesian Network Construction for Pediatric Patients on ECMO." AIME (2025).

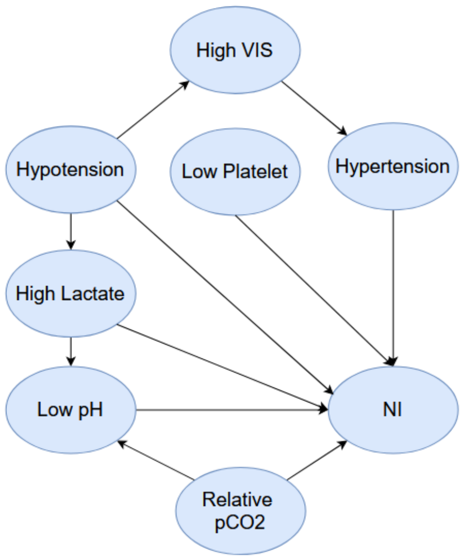
Mooney and Shavlik. "A Recap of Early Work on Theory and Knowledge Refinement." AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (2021).

Full causal graph + Refinement

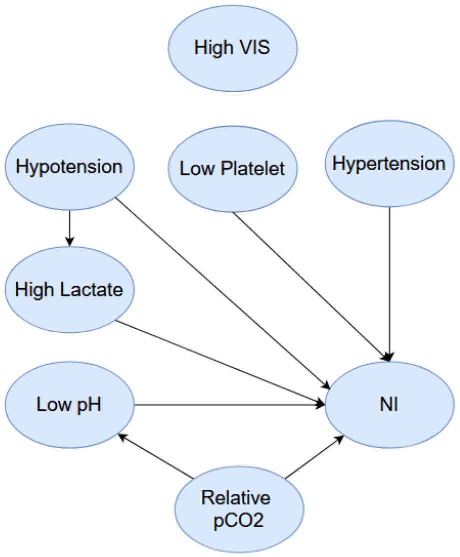
Clinician

Make a causal graph for ECMO, using these variables [...]

Claude



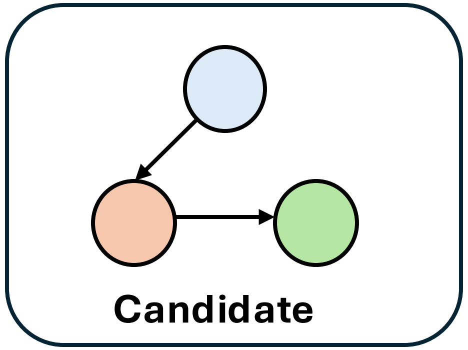
Refinement



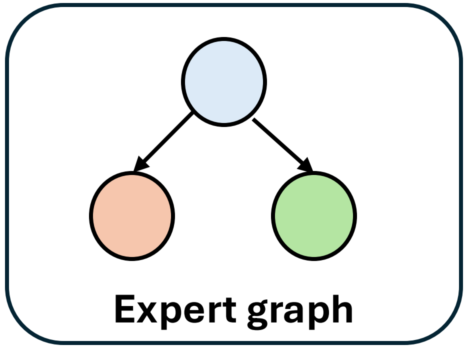
Reasoning

LLMs can't truly reason and can form spurious associations

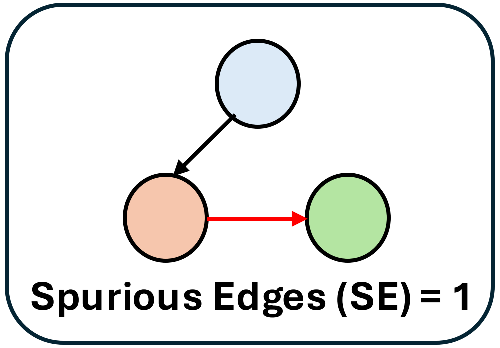
Evaluation scheme

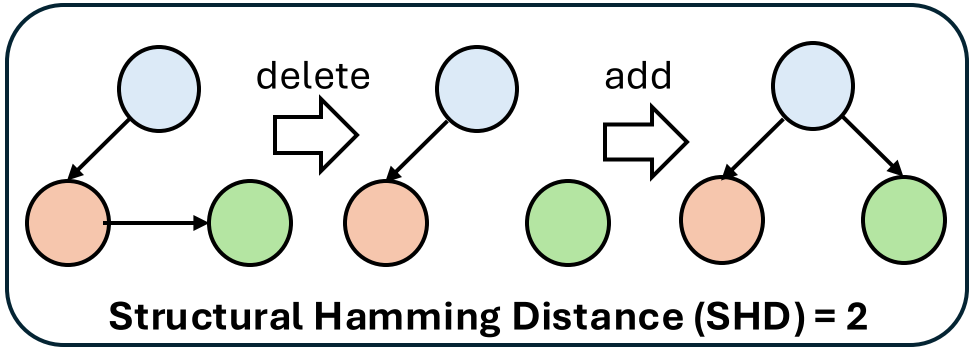


Generated by
1. Asking LLM about each edge separately
2. Directly eliciting full graph from LLM
3. Refining LLM's graph using data & expert knowledge



Metrics:





Quantitative Results:

LLMs perform poorly for Causal Question Answering

Study	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
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

During pairwise causal Question Answering, LLMs tend to generate contradictory answers leading to higher edge deletion

Quantitative Results:

LLMs perform poorly for Causal Question Answering

PELICAN

Method	Metric SHD	SID	SE
Fast Causal Inference	8.0 ± 0.5	14.9 ± 0.3	0.1 ± 0.3
Claude (Pairwise)	6	6	1
GPT 4o (Pairwise)	9	14	6
OpenBioLLM (Pairwise)	23	11	22

nuMoM2b

Method	Metric SHD	SID	SE
Fast Causal Inference	31.8 ± 1.2	80.6 ± 4.1	2.3 ± 1.7
Claude (Pairwise)	17	44	9
GPT 4o (Pairwise)	20	53	10
OpenBioLLM (Pairwise)	32	54	22

Pairwise querying across many LLMs may at times be as bad or worse than data-driven causal discovery

LLM-generated causal graphs better than data-driven causal discovery, but still quite far away from expert graphs

Quantitative Results:

LLMs might act as approximate knowledge bases

PELICAN

Method	Metric		
	SHD	SID	SE
Fast Causal Inference	8.0 ± 0.5	14.9 ± 0.3	0.1 ± 0.3
Claude (Full)	4	5	4
GPT 4o (Full)	8	15	6
OpenBioLLM (Full)	9	19	5
Claude (Full + Refine)	4.5 ± 0.7	5 ±1.2	2.1 ± 0.3
GPT 4o (Full + Refine)	7 ± 1	12 ± 3.7	1.6 ± 0.9
OpenBioLLM (Full + Refine)	7.8 ± 1.2	14.3 ± 1.5	0.9 ± 0.9

nuMoM2b

Method	Metric		
	SHD	SID	SE
Fast Causal Inference	31.8 ± 1.2	80.6 ± 4.1	2.3 ± 1.7
Claude (Full)	17	44	9
GPT 4o (Full)	20	53	5
OpenBioLLM (Full)	32	54	22
Claude (Full + Refine)	18.5 ± 0.8	49.5 ± 0.8	8.2 ± 0.6
GPT 4o (Full + Refine)	21.9 ± 0.8	58.5 ± 3.4	4.1 ± 0.3
OpenBioLLM (Full + Refine)	26.2 ± 1.0	45 ± 4.4	13.5 ± 0.5

Refinement improves SE & SHD

Conclusion & Future Work

Conclusion

- LLMs being used by clinicians
- Evaluate LLMs as AI-in-the-loop
- Results on two domains
  - Bad at pairwise causal QA,
  - Unreliable as exact Knowledge Bases
  - Better as approximate Knowledge Bases

Future Work

- Additional Domains
- Relation b/w domain characteristics & LLM performance
- Possible improvements to setup
  - Validity constraints
  - Retrieval Augmented Generation
  - LLM ensembles