Causal Discovery with Language Models as Imperfect Experts

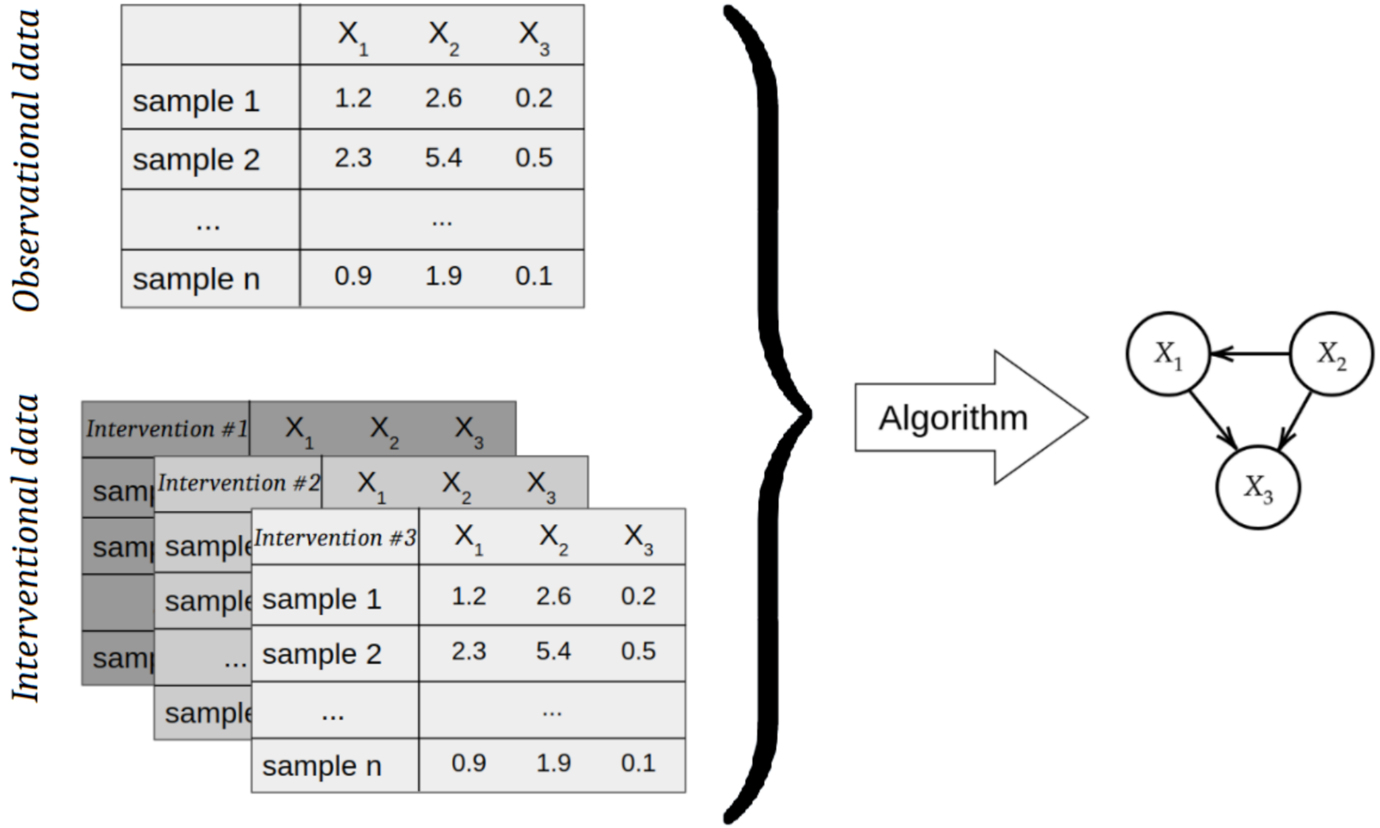
Stephanie Long [*1] Alexandre Piché [*234] Valentina Zantedeschi [*2] Tibor Schuster [1] Alexandre Drouin [24]

Causality Discussion Group - 2024

Scope of the work

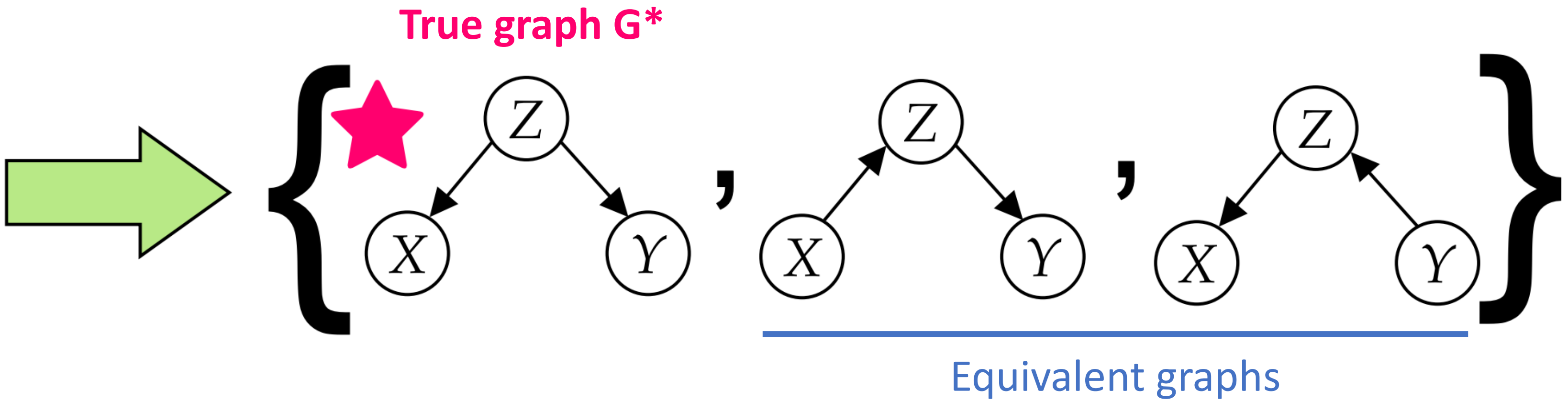
- NOT ABOUT causal reasoning of Large Language Models
- ABOUT leveraging information from related tasks
  - by querying an (imperfect) expert
  - via variables' meta-data (e.g., their name or description)
  - to reduce uncertainty in data-based causal discovery methods

Causal Discovery



Markov Equivalence Class

X	Y	Z
1.21	1.58	0.33
1.50	1.84	0.51
⋮	⋮	⋮
0.96	1.07	0.11

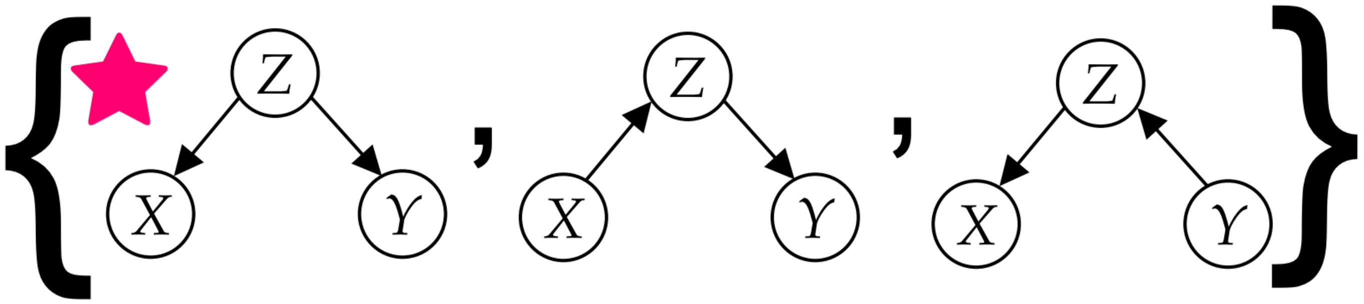


How to reduce uncertainty?

Causal Discovery with Expert Knowledge

X	Y	Z
1.21	1.58	0.33
1.50	1.84	0.51
⋮	⋮	⋮
0.96	1.07	0.11

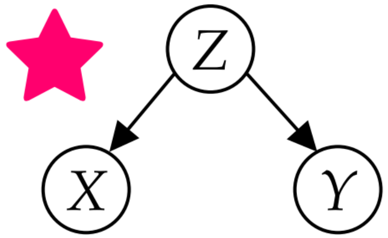
Data-driven
Algorithm
e.g., PC



$p(Z \rightarrow Y) = 0.9$ $p(Z \leftarrow Y) = 0.1$
$p(Z \rightarrow X) = 0.5$ $p(Z \leftarrow X) = 0.5$

Expert orientations and their probability

Bayesian Inference



Key point: We do NOT assume that the experts are perfect

Causal Discovery with Expert Knowledge

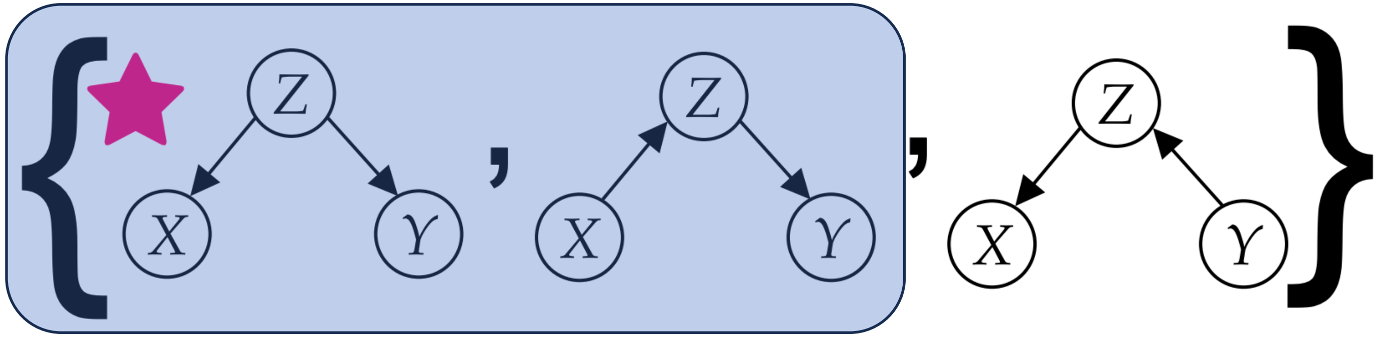
$$\min \ \left|\mathcal{M}^{E,S}\right|$$

Final MEC size

$$\text{such that } p\left(G^{\star} \in \mathcal{M}^{E,S}\right) \geq 1 - \eta,$$

Probability orientations are correct

Tolerance to error



Causal Discovery with Expert Knowledge

$$\min \left|\mathcal{M}^{E,S}\right|$$

Final MEC size

$$\text{such that } p\left(G^{\star} \in \mathcal{M}^{E,S}\right) \geq 1 - \eta,$$

Probability orientations are correct

Tolerance to error

estimated via Bayesian inference:

P(edges are correctly oriented | we observed such expert orientations)

Causal Discovery with Expert Knowledge

Final MEC size

$$\min \left|\mathcal{M}^{E,S}\right|$$

$$\text{such that } p(G^\star \in \mathcal{M}^{E,S}) \geq 1-\eta,$$

Probability orientations are correct

Tolerance to error

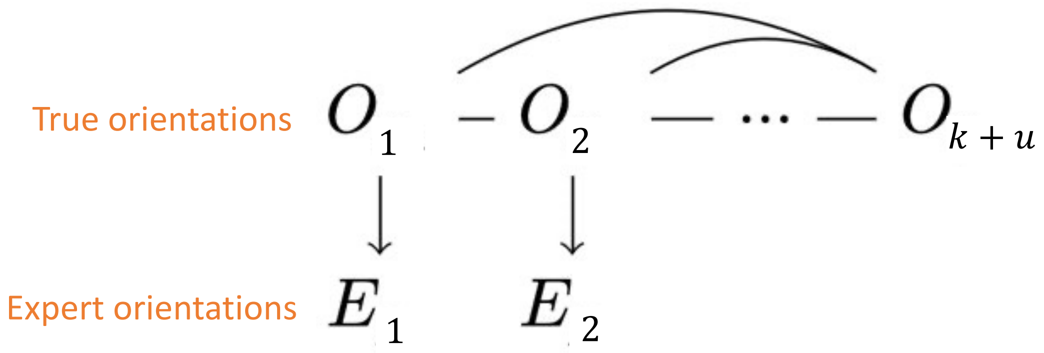
Hyper-parameter

Expert Model

Assumption: Expert makes independent decisions

$$p(E_1, \dots, E_u \mid O_1, \dots, O_u, \dots, O_{k+u}) = \prod_i p(E_i \mid O_i)$$

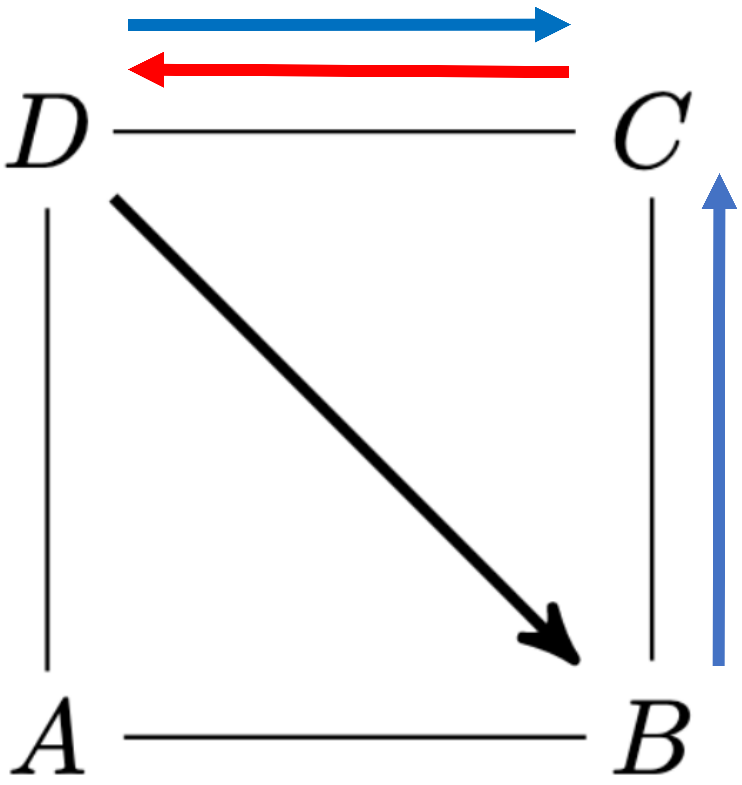
We can factorize the likelihood

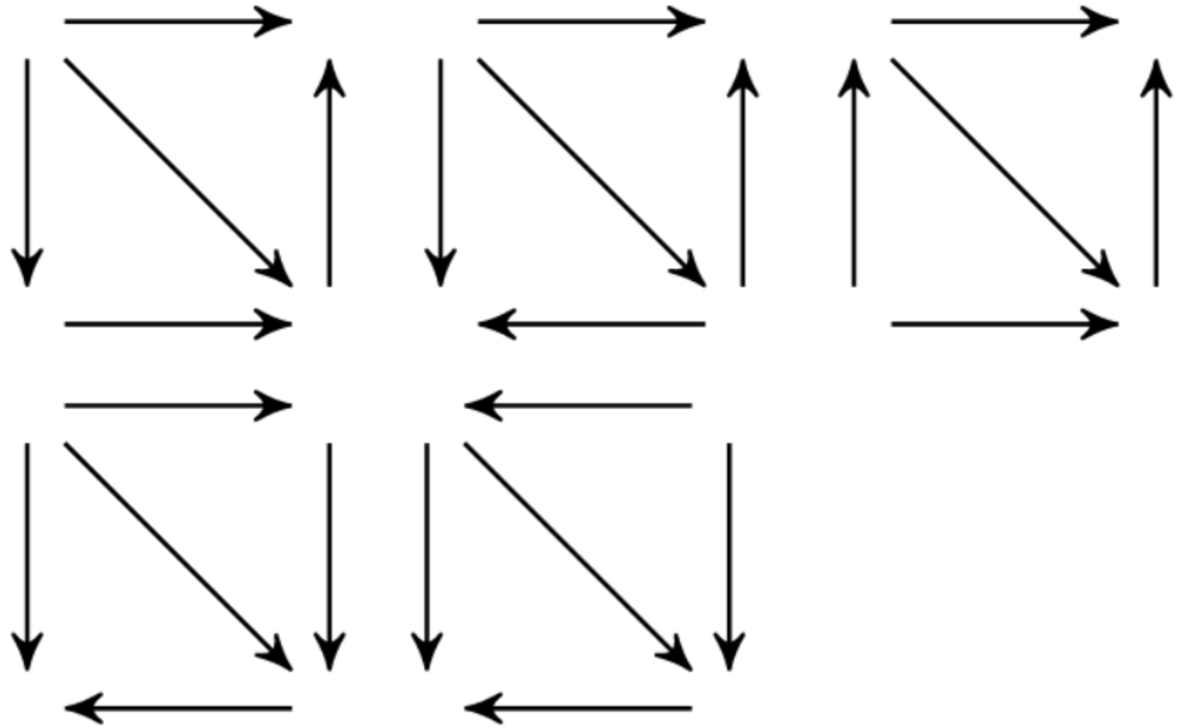


Bayesian Posterior

$$\underbrace{p(O_1, O_2 \mid E_1, \ldots, E_u)}_{\text{posterior}} = \frac{\overbrace{p(E_1,\ldots,E_u \mid O_1,O_2)}^{\text{Likelihood}}\ \overbrace{p(O_1,O_2)}^{\text{Prior}}}{\underbrace{p(E_1,\ldots,E_u)}_{\text{Normalization constant}}}$$

Edge orientations are inter-dependent

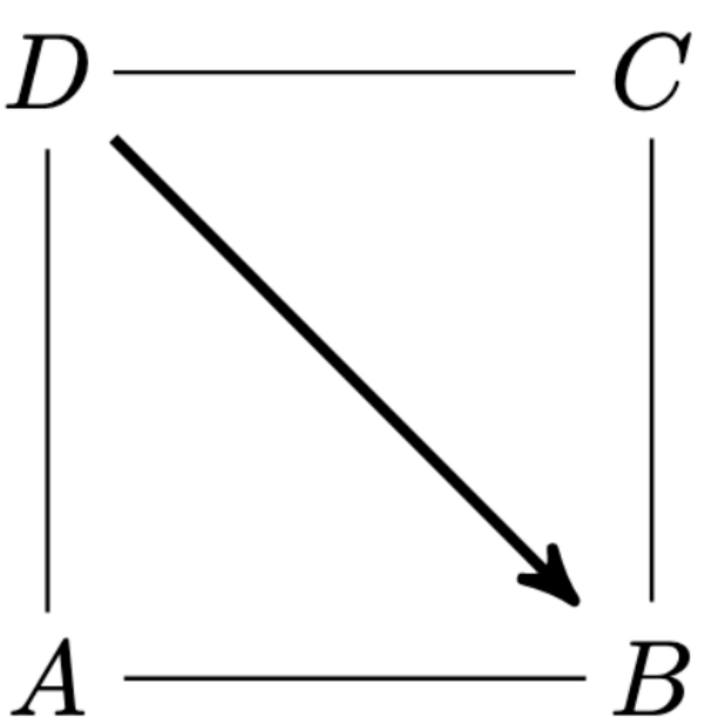


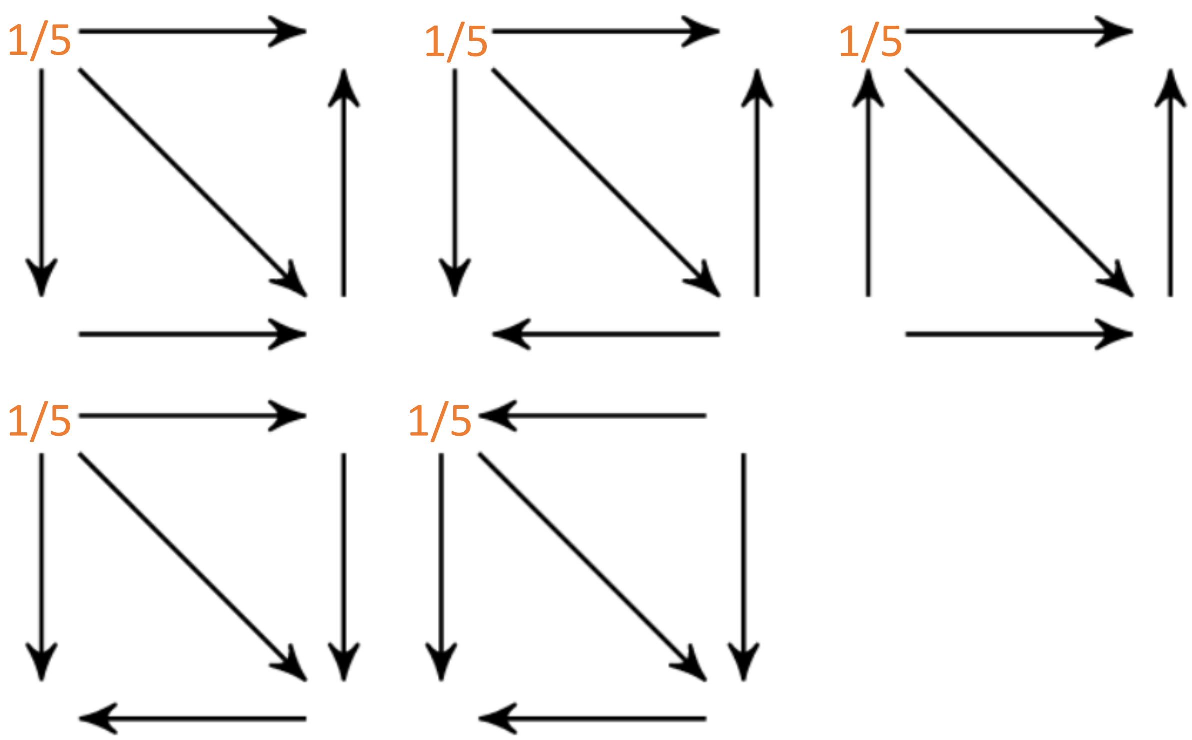


Posterior cannot be factorized as we do for the likelihood

Perković, Emilija, Markus Kalisch, and Maloes H. Maathuis. "Interpreting and using CPDAGs with background knowledge." arXiv preprint arXiv:1707.02171 (2017).

Edge orientations are inter-dependent





Prior := uniform over graphs in MEC

(we marginalize to get prior and posterior probabilities of subset of edges)

Perković, Emilija, Markus Kalisch, and Maloes H. Maathuis. "Interpreting and using CPDAGs with background knowledge." arXiv preprint arXiv:1707.02171 (2017).

Considered experts

- ϵ-expert: gives wrong orientation with constant probability of error
- LLM: ? we trust their confidence estimate

Are LLMs calibrated?

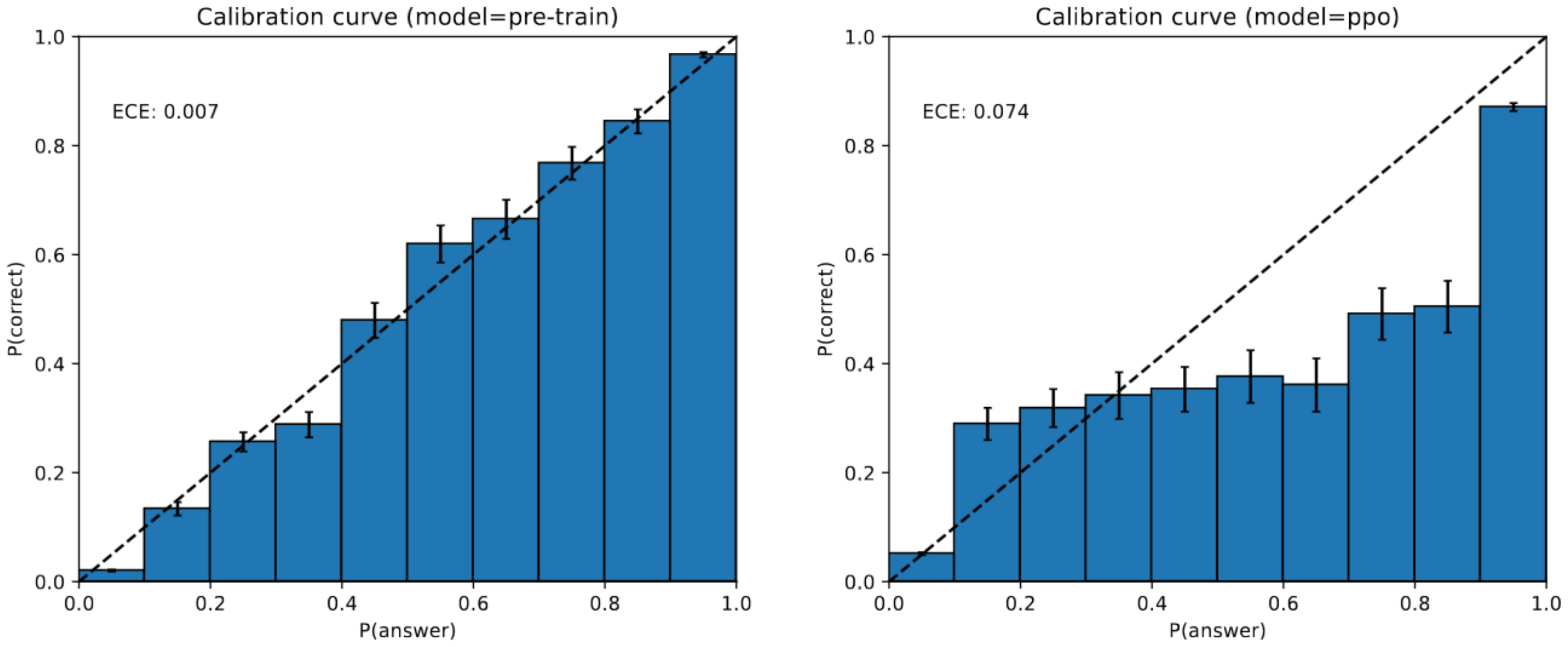


Figure 8. Left: Calibration plot of the pre-trained GPT-4 model on a subset of the MMLU dataset. On the x-axis are bins according to the model's confidence (logprob) in each of the A/B/C/D choices for each question; on the y-axis is the accuracy within each bin. The dotted diagonal line represents perfect calibration. Right: Calibration plot of the post-trained GPT-4 model on the same subset of MMLU. The post-training hurts calibration significantly.

Achiam, Josh, et al. "Gpt-4 technical report." *arXiv preprint arXiv:2303.08774* (2023).
Kadavath, Saurav, et al. "Language models (mostly) know what they know." *arXiv preprint arXiv:2207.05221* (2022).

Scoring orientations with LLMs

```
Among these two options which one is the most likely true:
(A) lung cancer causes cigarette smoking
(B) cigarette smoking causes lung cancer'
The answer is:
```

We compute likelihood of (A) and likelihood of (B)
… and normalize

Randomizing the prompt

```
Among these two options which one is the most likely true:
(A) {μ_i} {verb_k} {μ_j}
(B) {μ_j} {verb_k} {μ_i}
The answer is:
```

Greedy Algorithm

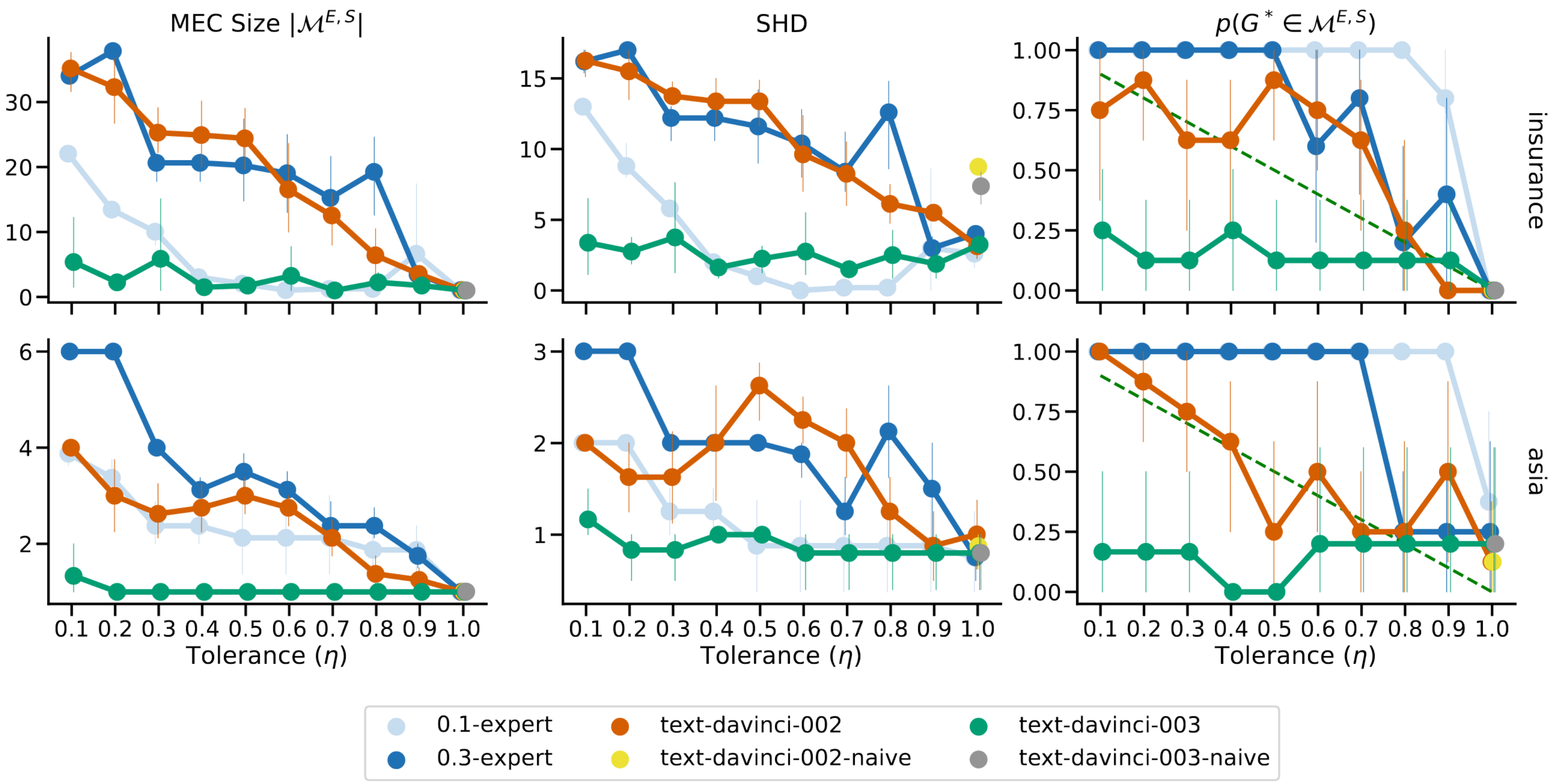
$$\min \left|\mathcal{M}^{E,S}\right|$$
$$\text{such that } p\left(G^{\star} \in \mathcal{M}^{E,S}\right) \geq 1-\eta$$

1. Query expert on all unoriented edges $(E_1, \dots, E_u)$
2. FOR each potential new orientation O_i, we compute the posterior:

   $$p(O_i, O_I \mid E_1, \dots, E_u)$$

   Where O_I is the set of orientations consequential to orienting O_i
3. Select (o_i, o_I) with the highest posterior
4. IF posterior of updated graph does not satisfy tolerance constraint, STOP
5. ELSE back to 2.

Results



MEC Size $|\mathcal{M}^{E,S}|$
SHD
$p(G^* \in \mathcal{M}^{E,S})$
insurance
asia
Tolerance (η)
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
0.1-expert
0.3-expert
text-davinci-002
text-davinci-002-naive
text-davinci-003
text-davinci-003-naive

Future Work

- Expert model is quite unrealistic

  How to account for systematic errors?

- Computing posterior requires enumerating all graphs in MEC

  How to scale to large number of variables?

Thanks!

- https://arxiv.org/abs/2307.02390
- https://github.com/StephLong614/Causal-disco-LLM-imperfect-experts