Graphical Perspectives on Causal Reasoning

Jacqueline Maasch ✦ Cornell Tech Computer Science

3^{rd} Learning on Graphs Workshop ✦ Flatiron Institute ✦ New York, New York ✦ 22 April 2025

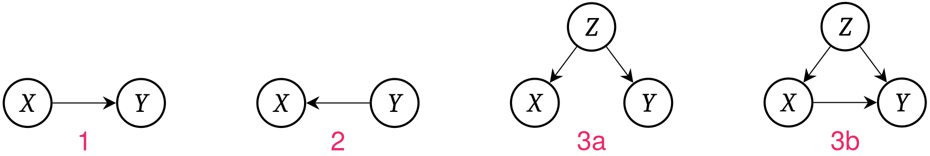
Causal Graphical Modeling

✦

Association $\neq$ causation

Reichenbach's Common Cause Principle [1]: Statistical association can be explained by

1. X is a cause of Y;
2. Y is a cause of X; or
3. X and Y are both caused by a third variable, **confounder** Z.

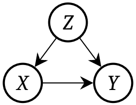


Structural causal models

1 Preliminaries: Causal Graphical Modeling

A **structural causal model** (SCM) [2] is a tuple $\mathcal{M} \coloneqq \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, p(\mathbf{u}) \rangle$:

- $\mathbf{U} = \{U_i\}_{i=1}^n$ are exogenous variables determined by factors outside $\mathcal{M}$;
- $\mathbf{V} = \{V_i\}_{i=1}^n$ are observed endogenous variables determined by variables in $\mathbf{U} \cup \mathbf{V}$;
- $\mathcal{F} = \{f_i\}_{i=1}^n$ are structural functions such that $v_i = f_i(\mathbf{pa}_{v_i}, u_i)$;
- $p(\mathbf{u})$ is the distribution over $\mathbf{U}$.



SCMs are associated with graphical representations, often **directed acyclic graphs** (DAGs).

Structural causal models

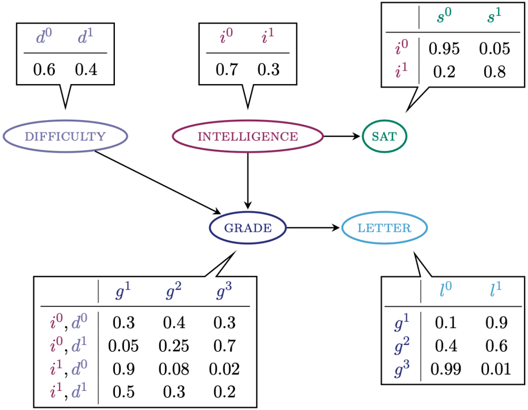
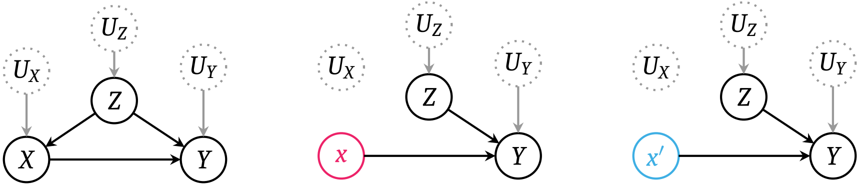


Figure adapted from [3].

Causal submodels



SCM $\mathcal{M}$ Submodel $\mathcal{M}_x$ Submodel $\mathcal{M}_{x'}$

Interventions $do(x)$, $do(x')$ replace the generative mechanism for X with a constant function.

Figure adapted from [4]

Reasoning in AI

What is reasoning?

Many **general** definitions exist, e.g.:

- Logical process of drawing valid conclusions from new information and prior knowledge.
- Bayesian inference: $P(H \mid E) = \frac{P(E|H)P(H)}{P(E)}$, where H is hypothesis and E is evidence.

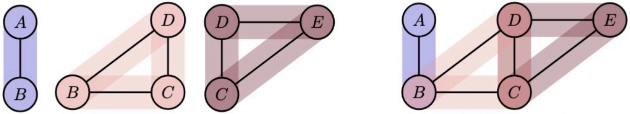
Many **specific** forms: quantitative, logical, visual, spatial, moral, legal, etc.

We consider two forms in tandem:

1. **Causal**: reasoning about cause and effect in factual and counterfactual worlds.
2. **Compositional**: recognizing + synthesizing novel combos of previously seen concepts.

Compositionality + causality

- **Graphical modeling.** Expressive representations for joint distributions, their factors, and the propagation of quantities through systems [5, 6, 7].
- **Causal inference.** Causal effect decomposition in mediation analysis [8, 9], fairness analysis [10], covariate adjustment with latent variables [11, 12], etc.



Local subgraphs. *Global graph.*

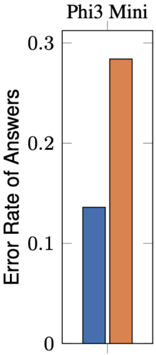
The recall-reasoning gap

Factual Question

A number is divisible by six if it has both two and three as prime factors. Is $\{N\}$ divisible by six?

Counterfactual Question

A number is divisible by six if it has both two and three as prime factors. Suppose that $\{N\}$ had three as one of its prime factors (retaining all its other prime factors). Then, would it have been been divisible by six?



Lower error rate on factual questions (**recall**) than counterfactual questions (**reasoning**).[1]

[1] Sampling 10 answers for each $N \in \{1, \ldots, 100\}$ [13].

Two intertwined issues

2 Preliminaries: Reasoning in AI

How to improve reasoning?

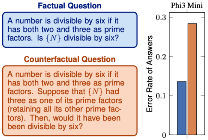
Published as a conference paper at ICLR 2025

REASONING ELICITATION IN LANGUAGE MODELS VIA COUNTERFACTUAL FEEDBACK

Alihan Hüyük,[*†] **Xinnuo Xu,**[‡] **Jacqueline Maasch,**[§] **Aditya V. Nori,**[‡] **Javier González**[‡]
[†]Harvard University, [‡]Microsoft Research Cambridge, [§]Cornell Tech

1 INTRODUCTION

Large language models (LLMs) are shown to be capable of delivering astounding performance in numerous tasks across various domains. Examples stretch from writing assistants (Gan et al., 2023), to sentiment analysis in social media (Simmering and Huoviala, 2023), and even applications in healthcare (González et al., 2023; Wong et al., 2023). While the ever-increasing accuracy of these systems is now undeniable, it is still rather unclear to what extent this accuracy is due to effective *recall* of their training data vs. a genuine ability to *reason* by extracting, understanding, and adapting the fundamental



ICLR '25

How to measure reasoning?

Compositional Causal Reasoning Evaluation in Language Models

Jacqueline R. M. A. Maasch[1] **Alihan Hüyük**[2] **Xinnuo Xu**[3] **Aditya V. Nori**[3] **Javier Gonzalez**[3]

Abstract

Causal reasoning and compositional reasoning are two core aspirations in generative AI. Measuring the extent of these behaviors requires principled evaluation methods. We explore a unified perspective that considers both behaviors simultaneously, termed *compositional causal reasoning* (CCR): the ability to infer how causal measures compose and, equivalently, how causal quantities propagate through graphs. We instantiate a framework for the systematic evaluation of CCR for the average treatment effect and the probability of necessity and sufficiency. As proof of concept, we demonstrate the design of CCR tasks for language models in the LLama, Phi, and GPT families. On a math word problem, our framework revealed a range of taxonomically distinct error patterns. Additionally, CCR errors increased with the complexity of causal paths for all models except o1.

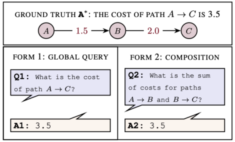


Figure 1. Compositionally consistent responses to two formulations of a simple (non-causal) query. Reasoning is externally valid if A1 and A2 both equal A*, and internally consistent if A1==A2.

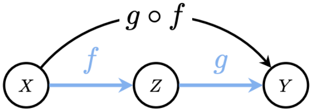
Baroni, 2023).[1] It is both a means of generalization and of coping with complexity: problems can be reformulated as simpler subproblems connected by compositional rules.

ICML '25

Compositional Causal Reasoning Evaluation in LMs

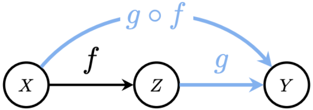
Compositional causal reasoning (CCR)

The ability to infer **compositions and decompositions** of causal measures in factual and counterfactual worlds.



A. INFER $g \circ f$ FROM f, g

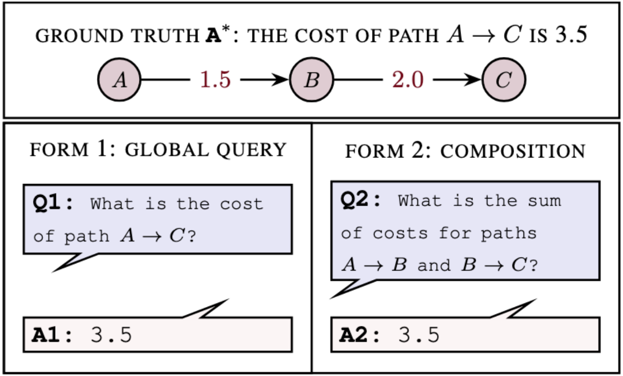
Inductive CCR



B. INFER f FROM $g \circ f$, g

Deductive CCR

A noncausal example for intuition

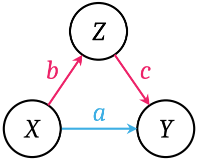


A classic causal example

3 Compositional Causal Reasoning Evaluation in LMs

Example 1: Decomposition of total causal effects in linear SCMs [2]

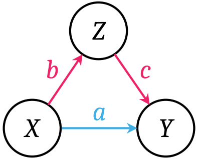
$$\underbrace{\mathrm{TE}}_{\textit{total effect}} = \underbrace{\mathrm{NDE}}_{\textit{direct effect}} + \underbrace{\mathrm{NIE}}_{\textit{indirect effect}} \tag{1}$$



$$\mathrm{TE}_{XY} = a + bc$$

Compositional consistency

Reasoning is **compositionally consistent** when theoretically equivalent compositions are inferred to be equal.



$$\mathrm{TE}_{XY} = a + bc$$

Example 1: a causal reasoning agent should infer that TE is equivalent to NDE + NIE.

External validity

Reasoning is **externally valid** when causal estimates are equivalent to ground truth, up to some error δ for error metric θ:

$$\theta(\underbrace{\varphi^*_{\mathbf{x}}}_{\text{true}}, \underbrace{\widehat{\varphi}_{\mathbf{x}}}_{\text{estimate}}) \leq \delta. \tag{2}$$

From Example 1:

- $\theta(\mathrm{TE}^*_{XY}, \widehat{\mathrm{TE}}_{XY}) \leq \delta$
- $\theta(\mathrm{TE}^*_{XY}, \widehat{\mathrm{NDE}}_{XY} + \widehat{\mathrm{NIE}}_{XY}) \leq \delta$
- Etc.

Internal consistency

Reasoning is **internally consistent** when quantities that are theoretically equivalent are inferred to be equivalent, up to some error δ:

$$\underbrace{\varphi^*_{\mathbf{x}} = \varphi^*_{\mathbf{x}'}}_{\text{equal in truth}} \Rightarrow \underbrace{\theta(\widehat{\varphi}_{\mathbf{x}}, \widehat{\varphi}_{\mathbf{x}'}) \leq \delta.}_{\text{equal in estimation}} \tag{3}$$

Estimates are compared to each other, not to ground truth. From Example 1:

- $\theta(\widehat{\mathrm{TE}}_{XY}, \widehat{NDE}_{XY} + \widehat{NIE}_{XY}) \leq \delta$

Taxonomy of reasoners

External validity		
	Valid-consistent (VC)	Valid-inconsistent (VI)
	Invalid-consistent (IC)	Invalid-inconsistent (II)
	Internal consistency	

Probability of necessity and sufficiency (PNS)

Let X and Y denote binary random variables, where X is a cause of Y. The probability that event x ($X =$ *true*) is necessary and sufficient to produce event y ($Y =$ *true*) is [14]

$$\text{PNS} := \mathbb{P}(y_x, y'_{x'}). \tag{4}$$

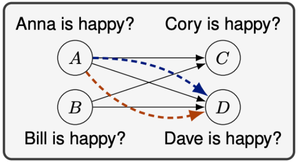
When Y is **monotonic** in X, this is identifiable by the following expression:

$$\text{PNS} = \mathbb{P}(y_x) - \mathbb{P}(y_{x'}) = \mathbb{P}(y \mid do(x)) - \mathbb{P}(y \mid do(x')) = \text{ATE}. \tag{5}$$

CandyParty task

3 Compositional Causal Reasoning Evaluation in LMs

Question: Anna, Bill, Cory, and Dave are going to a party, where the host is going to distribute candies. Anna will be happy if she gets at least 4 candies. Bill will be happy if he gets at least 6 candies. Cory will be happy if Anna and Bill are both happy or if he gets at least 8 candies. Dave will be happy if Anna and Bill are both happy or if he gets at least 10 candies. After distributing the candies, Anna gets N_A, Bill gets N_B, Cory gets N_C, and Dave gets N_D. Is Dave happy?



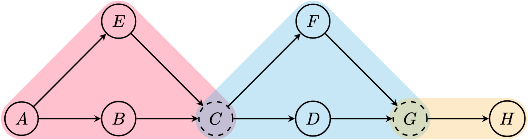
$$N_A, N_B, N_C, N_D \sim \mathcal{U}(1, \ldots, 12)$$
$$A = N_A \geq 4$$
$$B = N_B \geq 6$$
$$C = (A \wedge B) \vee (N_C \geq 8)$$
$$D = (A \wedge B) \vee (N_D \geq 10)$$

Counterfactual Question: Now, suppose that Anna is (not) happy regardless of the candy distribution. With this assumption, is Dave happy?

True/false questions, logical operators (and, or),

Probability of necessity and sufficiency (PNS)

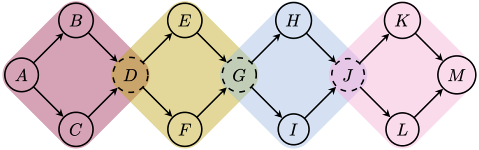
Multiplicative composition across biconnected components (BCCs):



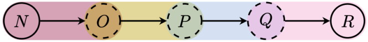
$$\mathrm{PNS}_{AH} = \mathrm{PNS}_{AC} \cdot \mathrm{PNS}_{CG} \cdot \mathrm{PNS}_{GH}$$

Commutative cut trees (CCTs)

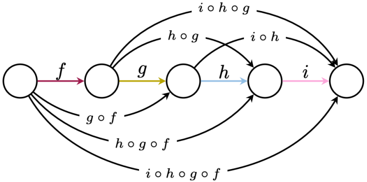
3 Compositional Causal Reasoning Evaluation in LMs



A. DAG $\mathcal{G}_{AM}$



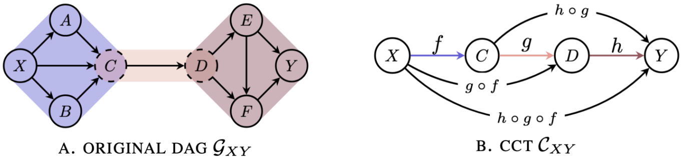
B. DAG $\mathcal{G}_{NR}$



C. CCT

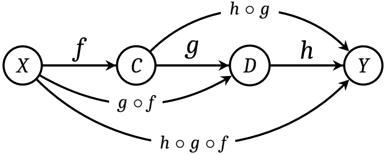
Running example

3 Compositional Causal Reasoning Evaluation in LMs



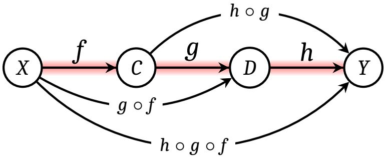
Global	PNS_{XY}
Local	PNS_{XC}, PNS_{XD}, PNS_{CD}, PNS_{CY}, PNS_{DY}
Composition	$\mathrm{PNS}_{XC}\mathrm{PNS}_{CY}$, $\mathrm{PNS}_{XD}\mathrm{PNS}_{DY}$, $\mathrm{PNS}_{XC}\mathrm{PNS}_{CD}\mathrm{PNS}_{DY}$

Pathways of reasoning



Note: $f := \mathrm{PNS}_{XC}$, $g := \mathrm{PNS}_{CD}$, $h := \mathrm{PNS}_{DY}$, $g \circ f := \mathrm{PNS}_{XD}$, $h \circ g \circ f := \mathrm{PNS}_{XY}$, etc., and composition is multiplicative.

Pathways of reasoning

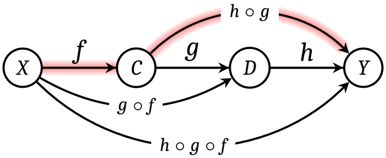


$$\mathrm{PNS}_{XC}\mathrm{PNS}_{CD}\mathrm{PNS}_{DY}$$

Note: $f := \mathrm{PNS}_{XC}$, $g := \mathrm{PNS}_{CD}$, $h := \mathrm{PNS}_{DY}$, $g \circ f := \mathrm{PNS}_{XD}$, $h \circ g \circ f := \mathrm{PNS}_{XY}$, etc., and composition is multiplicative.

Pathways of reasoning

3 Compositional Causal Reasoning Evaluation in LMs

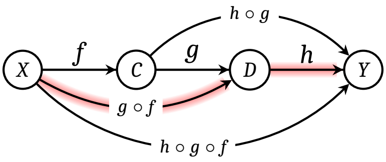


$$\mathrm{PNS}_{XC}\mathrm{PNS}_{CY}$$

Note: $f := \mathrm{PNS}_{XC}$, $g := \mathrm{PNS}_{CD}$, $h := \mathrm{PNS}_{DY}$, $g \circ f := \mathrm{PNS}_{XD}$, $h \circ g \circ f := \mathrm{PNS}_{XY}$, etc., and composition is multiplicative.

Pathways of reasoning

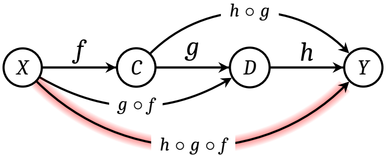
3 Compositional Causal Reasoning Evaluation in LMs



$$\mathrm{PNS}_{XD}\mathrm{PNS}_{DY}$$

Note: $f := \mathrm{PNS}_{XC}$, $g := \mathrm{PNS}_{CD}$, $h := \mathrm{PNS}_{DY}$, $g \circ f := \mathrm{PNS}_{XD}$, $h \circ g \circ f := \mathrm{PNS}_{XY}$, etc., and composition is multiplicative.

Pathways of reasoning
3 Compositional Causal Reasoning Evaluation in LMs



PNS_{XY}

Note: $f := \mathrm{PNS}_{XC}$, $g := \mathrm{PNS}_{CD}$, $h := \mathrm{PNS}_{DY}$, $g \circ f := \mathrm{PNS}_{XD}$, $h \circ g \circ f := \mathrm{PNS}_{XY}$, etc., and composition is multiplicative.

LMs as counterfactual data simulators

- 1000 sets of exogenous variable values sampled per quantity of interest.
- One factual, one counterfactual problem per set. Five answers sampled per problem.
- **Valid estimates:** $\geq 90\%$ of estimates with relative absolute error (RAE) ≤ 0.1.
- **Near-valid estimates:** $\geq 75\%$ of estimates with RAE ≤ 0.1.

Three layers of evaluation

3 Compositional Causal Reasoning Evaluation in LMs

Factual prompt

```
Xinyu, Ara, Becca, Celine, Daphne, Emma, Fox,
and Yasmin are going to a party, where the host
is going to distribute candies.  Xinyu will
be happy if she gets at least 7 candies.  Ara
will be happy if Xinyu is happy or if he gets
at least 7 candies.  Becca will be happy if...
After distributing the candies, Xinyu gets 4,
Ara gets 6, Becca gets 5, Celine gets 10, Daphne
gets 1, Emma gets 1, Fox gets 4, and Yasmin gets
3.  Is Celine happy?  Be as concise as possible.
```

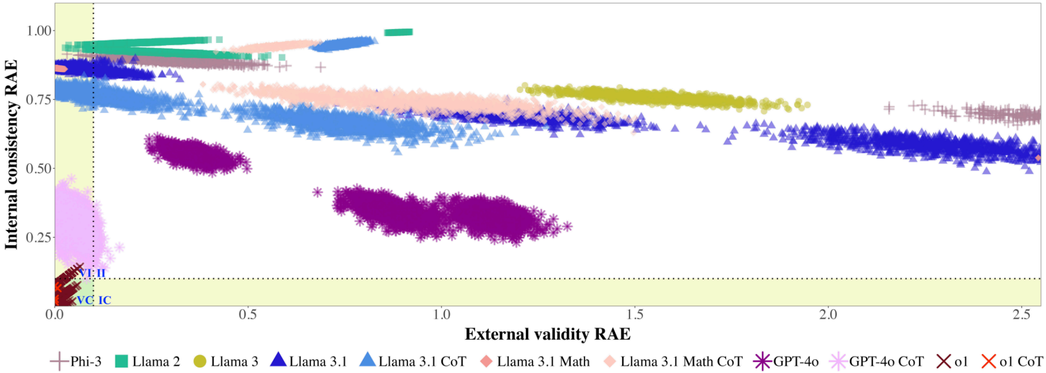
Counterfactual prompt

```
Now, suppose that Xinyu is happy regardless of
the candy distribution.  With this assumption,
is Celine happy?  Be as concise as possible.
```

True (factual)	Response (factual)	True (counterfactual)	Response (counterfactual)
True	True	True	False
True	True	True	False
True	True	True	False
True	True	True	False
True	True	True	False
...	...	...	...

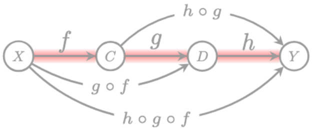
	Estimates derived from response vectors	
Associational	**Causal reasoning**	**CCR**
Precision, recall (etc.) of boolean responses	External validity per causal quantity	+ Internal consistency of causal compositions

Taxonomy of reasoners

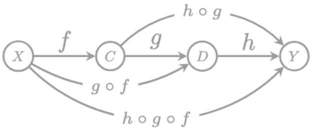


Composition RAE with respect to ground truth (**external validity**) and $\widehat{\mathrm{PNS}}_{XY}$ (**internal consistency**).
Dotted lines are error thresholds (RAE = 0.1).

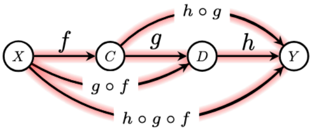
Visualizing reasoning errors with CCTs



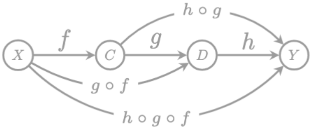
A. LLAMA 3.1 MATH



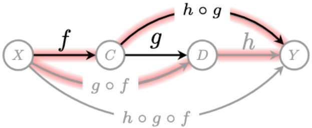
B. GPT-4O



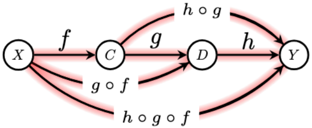
C. O1



D. LLAMA 3.1 MATH COT



E. GPT-4O COT



F. O1 COT

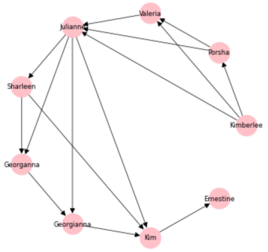
Error analysis

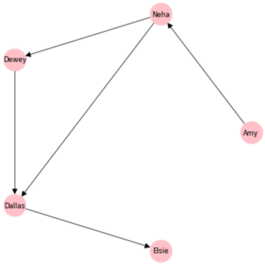
- **Some failure modes:**
  1. Failure to correctly extract causal relations.
  2. Incorrect logic despite correct causal relation extraction.
  3. Truncated reasoning process.
  4. Poor numeracy.

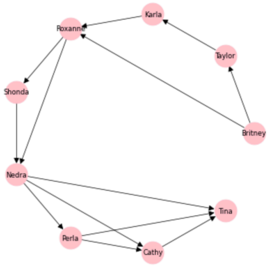
- **GPT-4o with CoT**: compositionally inconsistent despite local relation extraction.

- **VI reasoners** can fail to compose over multiple strands of logic even when correctly recapitulating relations directly expressed in the context prompt.
  $\Rightarrow$ *akin to passing a math quiz by memorizing specific answers, instead of synthesizing*.

Benchmark dataset 🤗 + random task generator







Thank you! Any questions?

maasch@cs.cornell.edu ✦ https://jmaasch.github.io/ ✦ arXiv:2503.04556

References

[1] H. Reichenbach. *The Direction of Time*. University of California Press, 1956.

[2] J. Pearl. "Direct and Indirect Effects". In: *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*. San Francisco. 2001.

[3] D. Koller et al. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[4] A. Darwiche. "Causal inference using tractable circuits". In: *arXiv preprint arXiv:2202.02891* (2022).

[5] J. Pearl. "Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach". In: *Proceedings, AAAI-82* (1982), pp. 133–136.

[6] G. R. Shafer et al. "Probability Propagation". In: *Annals of Mathematics and Artificial Intelligence* 2 (1990), pp. 327–351.

[7] F. R. Kschischang et al. "Factor graphs and the sum-product algorithm". In: *IEEE Transactions on information theory* 47.2 (2001), pp. 498–519.

[8] J. Pearl. "Interpretation and identification of causal mediation.". In: *Psychological methods* 19.4 (2014), p. 459.

[9] T. J. VanderWeele. "Mediation analysis: a practitioner's guide". In: *Annual review of public health* 37 (2016), pp. 17–32.

[10] D. Plečko et al. "Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning". In: *Foundations and Trends® in Machine Learning* 17.3 (2024), pp. 304–589. DOI: 10.1561/2200000106.

[11] J. Pearl. "Causal diagrams for empirical research". In: *Biometrika* 82.4 (1995), pp. 669–688.

References

[12] H. Jeong et al. “Finding and listing front-door adjustment sets”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 33173–33185.

[13] A. Hüyük et al. “Reasoning Elicitation in Language Models via Counterfactual Feedback”. In: *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. 2025.

[14] J. Pearl. “Probabilities of Causation: Three Counterfactual Interpretations and Their Identification”. In: *Synthese* 121 (1999), pp. 93–149.

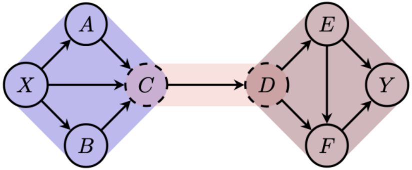
[15] J. Westbrook et al. “Maintaining bridge-connected and biconnected components on-line”. In: *Algorithmica* 7.1 (1992), pp. 433–464.

Models

Model	Parameters	Link
Phi-3-Mini-128K-Instruct (Abdin et al., 2024)	3.82B	https://huggingface.co/microsoft/Phi-3-mini-128k-instruct
Llama-2-7b-Chat-HF (Touvron et al., 2023)	6.74B	https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
Llama-3-8B-Instruct (Dubey et al., 2024)	8.03B	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Llama-3.1-8B-Instruct (Dubey et al., 2024)	8.03B	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
OpenMath2-Llama3.1-8B (Toshniwal et al., 2024)	8.03B	https://huggingface.co/nvidia/OpenMath2-Llama3.1-8B
GPT-4o	> 175B	https://openai.com/index/gpt-4o-system-card/
o1	> 175B	https://openai.com/o1/

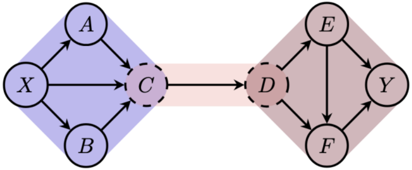
Table F.1. Large language models used for inference. The exact number of parameters in GPT-4o and o1 is not public knowledge, so we note the size of GPT-3 as a lower bound (B denotes billions).

Case Study: Graphs with Cutpoints



- A **cutpoint** is any node contained in multiple **biconnected components** (BCCs):
  - Maximal biconnected subgraphs induced by a partition of edges. Two edges are in the same partition if and only if they share a common simple cycle [15].
  - E.g., the blue, pink, and maroon subgraphs.
- Removing a cutpoint disconnects the graph (e.g., nodes C, D).

Assumptions: Graphs with Cutpoints



For simplicity, we consider causal DAGs satisfying the following:

- A1 Only one root node X (i.e., the cause of interest).
- A2 Only one leaf node Y (i.e., the effect of interest).
- A3 At least one cutpoint.
- A4 No unobserved confounders.

PNS composition across graph components

Theorem 1: PNS composition across biconnected components (BCCs)

Given DAG $\mathcal{G}_{XY}$ satisfying assumptions A1–A4 where Y is monotonic in X, the PNS for root X and leaf Y composes as

$$\mathrm{PNS}_{XY} = \prod_{\{R_i, L_i\} \in \mathbf{C}} \mathrm{PNS}_{R_i L_i} \tag{6}$$

where $\mathbf{C}$ is the set of all BCCs in $\mathcal{G}_{XY}$ and R_i, L_i are the root and leaf of BCC $\mathbf{C}_i$, respectively.

Factual and counterfactual prompts

Factual prompt

```
Xinyu, Ara, Becca, Celine, Daphne, Emma, Fox,
and Yasmin are going to a party, where the host
is going to distribute candies.  Xinyu will
be happy if she gets at least 7 candies.  Ara
will be happy if Xinyu is happy or if he gets
at least 7 candies.  Becca will be happy if...
After distributing the candies, Xinyu gets 4,
Ara gets 6, Becca gets 5, Celine gets 10, Daphne
gets 1, Emma gets 1, Fox gets 4, and Yasmin gets
3.  Is Celine happy?  Be as concise as possible.
```

Counterfactual prompt

```
Now, suppose that Xinyu is happy regardless of
the candy distribution.  With this assumption,
is Celine happy?  Be as concise as possible.
```

- $\widehat{\text{PNS}}_{XC}$: Simulate potential outcomes $X = \text{TRUE}$, $X = \text{FALSE}$ (Xinyu is or is not happy). Query for value of C (Celine is or is not happy).
- $\widehat{\text{PNS}}_{DY}$: Interventions on D (Daphne's happiness), queries on Y (Yasmin's happiness).
- **CoT formulation:** Demonstrated one factual and one counterfactual example.

Inductive CCR in Graphs with Cutpoints

Algorithm 1 *Inductive CCR evaluation in causal graphs with cutpoints*

Input: CCT $\mathcal{C}_{XY}$; estimates $\{\widehat{\varphi}\cdot\}$, true values $\{\varphi^*\cdot\}$ for $\langle\varphi, \mathcal{M}, \mathcal{Q}\rangle$; metric θ (e.g., relative absolute error)

Output: Reasoning errors η, ϵ, γ

Assumptions: φ composes according to an associative function over the BCCs of causal graph $\mathcal{G}_{XY}$.

Compute quantity-wise errors.

for $\forall$ pairs $\{R_i, L_{j>i}\}$ in $\mathcal{C}_{XY}$ **do**

  $\eta_{R_iL_j} \leftarrow \theta(\varphi^*_{R_iL_j}, \widehat{\varphi}_{R_iL_j})$ ▷ External validity.

Compute inductive reasoning errors.

for $\forall$ paths i from X to Y in $\mathcal{C}_{XY}$ **do**

  Get composition $\widehat{\varphi}_i^\circ$ for path i from knowledge of edges $j \in i$

  $\epsilon_i \leftarrow \theta(\varphi^*_{XY}, \widehat{\varphi}_i^\circ)$ ▷ External validity.

  $\gamma_i \leftarrow \theta(\widehat{\varphi}_{XY}, \widehat{\varphi}_i^\circ)$ ▷ Internal consistency.

return η, ϵ, γ

Commutative Cut Trees

Let $\mathcal{G}_{XY}$ be a causal graph satisfying A1–A4 and let φ be a causal measure that composes according to an associative function over BCCs (e.g., multiplication as in Theorem 1).

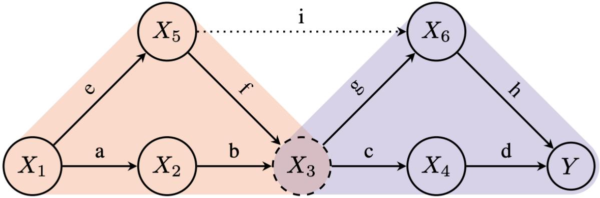
CCT $\mathcal{C}_{XY}$ is a transformation of $\mathcal{G}_{XY}$ that models all CCR pathways from root X to leaf Y for measure φ. $\mathcal{C}_{XY}$ is obtained by a two-step transformation of $\mathcal{G}_{XY}$:

1. Construct a causal chain with nodes $X \cup \mathbf{S} \cup Y$, where $\mathbf{S}$ is a topological ordering of the cutpoints in $\mathcal{G}_{XY}$.
2. Add a directed edge between any non-adjacent nodes in the chain to yield a complete graph where all directed paths point from root X to leaf Y.

CCTs: A useful abstraction

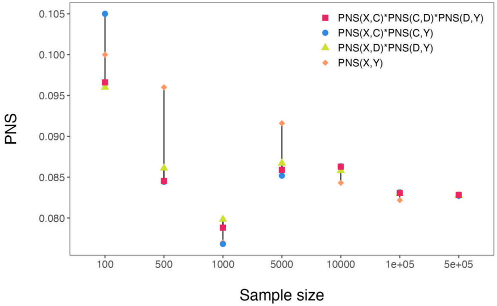
- Abstract away complexity in DAG by **collapsing BCCs** into single edges.
- Evaluate on complex DAGs with cutpoints as if they were **simply directed chains.**
- Simplify problem representation by (1) **marginalizing out variables** unnecessary for valid causal inference and (2) **visualizing pathways** of composition.
- A **design tool** for formulating reasoning tasks.
- Interpretable, intuitive tool for graphically **representing reasoning correctness**.

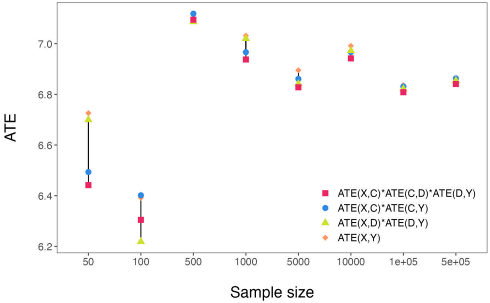
ATE Composition Across BCCs



- Assume a **linear SCM**.
- $\mathcal{G}_{X_1 Y}$ contains **subgraph with two BCCs** sharing cutpoint X_3 (in orange, periwinkle).
- If the dotted edge $X_5 \to X_6$ does not exist, $\mathrm{ATE}_{X_1 Y} = \mathrm{ATE}_{X_1 X_3} \cdot \mathrm{ATE}_{X_3 Y}$.
- If $X_5 \to X_6$ does exist, then product is summed with additional term corresponding to the path-specific effect for $X_1 \to X_5 \to X_6 \to Y$, which does not pass through X_3.

PNS & ATE Composition Across BCCs





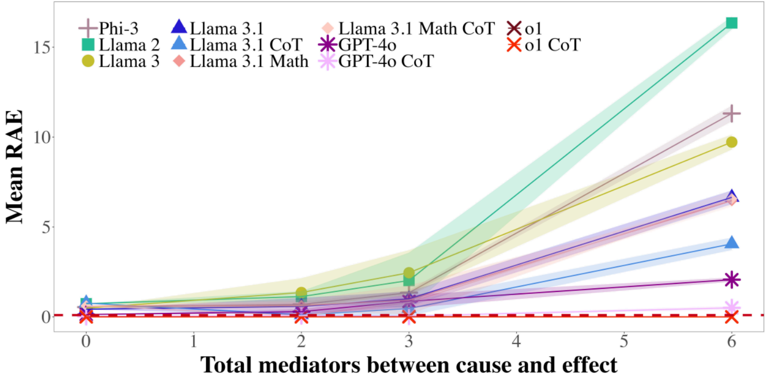
Extracting and evaluating PNS values

- 1000 sets of exogenous variable values sampled per quantity of interest.
- One factual, one counterfactual problem per set. Five answers sampled per problem.
- Responses converted to booleans using Llama 3 8B.
- **Approximation errors:** relative absolute errors (RAE).

$$RAE_{\text{external}} := \frac{|\, PNS^{*}\cdot - \widehat{PNS}\cdot \,|}{PNS^{*}\cdot}, \quad RAE_{\text{internal}} := \frac{|\, \widehat{PNS}\cdot - \widehat{PNS}\cdot' \,|}{\widehat{PNS}\cdot}. \tag{7}$$

- **Externally valid:** $\geq 90\%$ of estimates with RAE ≤ 0.1.
- **Near-valid:** $\geq 75\%$ of estimates with RAE ≤ 0.1.

Errors Increase With Mediation



Error Analysis

Response

```
Xinyu is not happy because she gets 6 candies, which is less than 7.  Ara is happy because he gets 8 candies, which is more than 7.  Celine is happy because Ara is happy.  **Daphne is not happy because she gets 2 candies, which is less than 7, and Celine is the only one who can make her happy.**  Emma is not happy because she gets 4 candies, which is less than 7, and Daphne is not happy.  Fox is not happy because he gets 3 candies, which is less than 7, and neither Daphne nor Emma is happy.  Yasmin is not happy because she gets 2 candies, which is less than 7, and neither Emma nor Fox is happy.  Therefore, Yasmin is not happy.
```

Figure F.8. A case where GPT-4o with CoT failed to reason correctly despite extracting the correct causal relations from a factual prompt used to compute $\widehat{\mathrm{PNS}}_{DY}$. The first reasoning error is in bold. As in Figure F.6, the model correctly implies that Celine (who is happy) is the only causal parent of Daphne in the underlying DAG, and yet still incorrectly concludes that Daphne is unhappy. This error then propagates through the remaining logic. This same error was observed multiple times from GPT-4o with CoT.

Error Analysis

Prompt

```
"After distributing the candies, Xinyu gets 4, Ara gets 6, Becca gets 5, Celine gets
10, Daphne gets 1, Emma gets 1, Fox gets 4, and Yasmin gets 3.  Is Celine happy?  Be
as concise as possible."
```

Response

```
"No, Celine is not happy.  She wants at least 7 candies but got only 10."
```

Figure F.3. Factual question and answer from Llama 2, displaying incorrect numerical reasoning.

Error Analysis

Prompt

```
"Now, suppose that Xinyu is happy regardless of the candy distribution.  With this
assumption, is Celine happy?  Be as concise as possible."
```

Response

```
"Under the assumption that Xinyu is happy regardless of the candy distribution,
Celine is not happy.

Reason:  Since Xinyu is happy regardless of the candy distribution, it means that
she got at least 7 candies.  However, Celine only got 10 candies, which is less than
7.  Therefore, Celine is not happy."
```

Figure F.4. Counterfactual question and answer from Llama 2, displaying incorrect numerical reasoning.

Error Analysis

Prompt

```
"After distributing the candies, Xinyu gets 10, Ara gets 6, Becca gets 7, Celine
gets 6, Daphne gets 1, Emma gets 10, Fox gets 10, and Yasmin gets 2.  Is Daphne
happy?  Be as concise as possible."
```

Response

```
"Daphne will be happy if Celine is happy or if she gets at least 7 candies.  Since
Daphne gets less than 7 candies (1), she won't be happy based on the candies she
gets.  Celine gets 6 candies, which is less than 7.  Since neither of Daphne's
conditions is met, Daphne is not happy."
```

Figure F.5. A case where Llama 3.1 failed to follow the full chain of logic. The correct line of logic is as follows: *Daphne will be happy if Celine is happy or if she gets at least 7 candies. Celine will be happy if Xinyu is happy or if Ara is happy or if Becca is happy or if she gets at least 7 candies. Xinyu will be happy if she gets at least 7 candies. Since Xinyu received 10 candies, she is happy, and so should be Celine. Thus, Daphne should also be happy.*