



# Compositional Causal Reasoning Evaluation in Language Models

Presenter: Jacqueline Maasch

[maasch@cs.cornell.edu](mailto:maasch@cs.cornell.edu) | arXiv:2503.04556

Jacqueline Maasch<sup>1</sup>, Alihan Hüyük<sup>2</sup>, Xinnuo Xu<sup>3</sup>, Aditya Nori<sup>3</sup>, Javier González<sup>3</sup>

<sup>1</sup>Cornell Tech, <sup>2</sup>Harvard University, <sup>3</sup>Microsoft Research Cambridge



# Presentation overview

## 1 Background

1. Background.
2. Compositional Causal Reasoning.
3. Compositional Consistency Evaluation.
4. Inductive CCR Evaluation in Graphs with Cutpoints.
5. Empirical Demonstration in Language Models.



# Reasoning in AI

## 1 Background

- There are many forms of reasoning: quantitative, logical, visual, spatial, moral, etc.
- We consider two forms in tandem: **causal reasoning** and **compositional reasoning**.
- Two major projects in AI reasoning:
  1. Increase reasoning capacities.
  2. Devise **principled evaluation frameworks**. ← Our work.



# Causal Reasoning

## 1 Background

- Causal reasoning is a defining outcome of human evolution [1].
- A causal lens enables humans and machines alike to learn generalizable lessons about the mechanics of the universe [2].
- Human-like AI might require reasoning at all 3 levels of Pearl's Causal Hierarchy [3]:
  1. Associational.
  2. Interventional.
  3. Counterfactual.



# Compositional Reasoning

## 1 Background

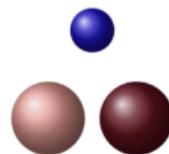
- Human-like AI might also require **compositional reasoning** [4]: the capacity to recognize and synthesize novel combinations of previously observed concepts [5].
- Compositionality is ubiquitous in the physical world, symbolic systems, human cognition [6], visual perception [7], and language [8].
- It is a means of **generalization** and of **coping with complexity**.

# Compositionality is Ubiquitous

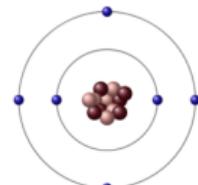
## 1 Background

*Algebra.* Let  $f(x) = 2x$ ,  $g(x) = x^2$ ,  $h(x) = x + 2$ . Solve  $h(g(f(x)))$ .

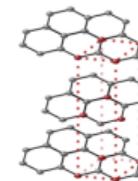
*Physics.*



Subatomic particles.



Carbon-12 atom.



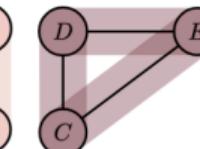
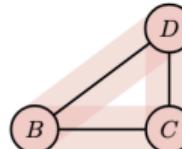
Graphite structure.

*Graphical modeling.*



A

Local subgraphs.



A

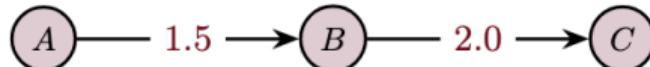
Global graph.



# Compositionality is Ubiquitous

## 1 Background

GROUND TRUTH **A\***: THE COST OF PATH  $A \rightarrow C$  IS 3.5



FORM 1: GLOBAL QUERY

**Q1:** What is the cost  
of path  $A \rightarrow C$ ?

**A1:** 3.5

FORM 2: COMPOSITION

**Q2:** What is the sum  
of costs for paths  
 $A \rightarrow B$  and  $B \rightarrow C$ ?

**A2:** 3.5



# Compositionality + Causality

## 1 Background

A mathematical language for evaluating compositional + causal reasoning simultaneously:

- **Graphical modeling.** Expressive representations for joint distributions, their factors, and the propagation of quantities through systems [9, 10, 11].
- **Causal inference.** Causal effect decomposition plays a central role in:
  - Mediation analysis [12, 13].
  - Fairness analysis [14].
  - Covariate adjustment in the presence of latent variables [15, 16].



# Contributions

## 2 Our Work

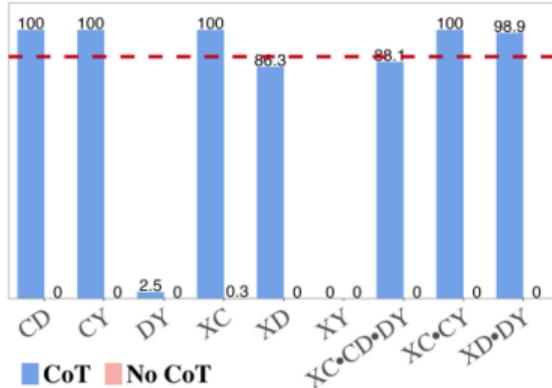
- 1. A compositional view of causal reasoning in LMs.** Formally express CCR as the ability of an LM to infer causal measure compositions (*inductive* reasoning) and decompositions (*deductive* reasoning).
- 2. Metrics and reasoning taxonomy.** Four categories of reasoners.
- 3. An evaluation framework.** For inductive CCR in causal graphs with cutpoints.
- 4. Preliminary empirical demonstration.** Evaluated CCR in seven LMs, with and without chain-of-thought (CoT) prompting.



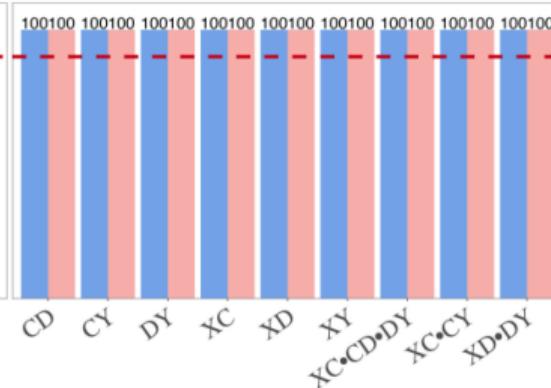
# Results Teaser

## 2 Our Work

GPT-4o



o1



- A simple CCR task revealed **taxonomically distinct error patterns**.
- Only **o1 was valid and consistent** on this task.



# Preliminaries: Causal Models

## 3 Preliminaries

**Definition 1: Structural causal model (SCM) [17]**

An SCM is a tuple  $\mathcal{M} := \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, p(\mathbf{u}) \rangle$ , where

- $\mathbf{U} = \{U_i\}_{i=1}^n$  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  $\mathbf{U}$ .



# Preliminaries: Causal Measures

## 3 Preliminaries

### Definition 2: Average treatment effect (ATE)

Let  $X$  denote a binary treatment variable and  $Y$  an outcome. We express the ATE as the following difference of expectations:

$$\text{ATE} := \mathbb{E}[Y \mid \text{do}(X = 1)] - \mathbb{E}[Y \mid \text{do}(X = 0)]. \quad (1)$$



# Preliminaries: Causal Measures

## 3 Preliminaries

In propositional logic, we say that

1.  $X$  is *necessary* for  $Y$  when  $Y \Rightarrow X$ .
2.  $X$  is *sufficient* for  $Y$  when  $X \Rightarrow Y$ .
3.  $X$  is *necessary and sufficient* for  $Y$  when  $X \iff Y$ .

Pearl [18] introduced a probabilistic framework, the **probabilities of causation (PrC)**:

1. Probability of necessity (PN).
2. Probability of sufficiency (PS).
3. Probability of necessity and sufficiency (PNS).



# Preliminaries: Causal Measures

## 3 Preliminaries

### Definition 3: Probability of necessity and sufficiency (PNS) [18]

Let  $X$  and  $Y$  denote binary random variables, where  $X$  is a cause of  $Y$ .

Let  $x$  and  $y$  denote the *propositions* or *events* that  $X = \text{TRUE}$  and  $Y = \text{TRUE}$ , respectively, while  $x'$  and  $y'$  denote that  $X = \text{FALSE}$  and  $Y = \text{FALSE}$ .

The probability that  $x$  is necessary and sufficient to produce  $y$  is given as

$$\text{PNS} := \mathbb{P}(y_x, y'_{x'}) = \mathbb{P}(x, y)\text{PN} + \mathbb{P}(x', y')\text{PS}. \quad (2)$$



# Preliminaries: Causal Measures

## 3 Preliminaries

Why use the PNS for reasoning evaluation?

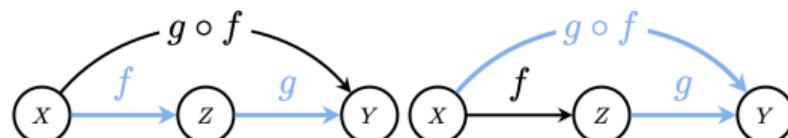
1. Variables of interest are **binary** and probabilities are **bounded by 0 and 1**.
2. Easy to translate **PrC queries to text prompts** designed to elicit logical, mathematical, probabilistic, and/or causal reasoning [19, 20].
3. The **PNS and ATE coincide** under certain conditions, and thus share convenient compositional properties.

# Compositional Causal Reasoning

## 4 Compositional Causal Reasoning

### Definition 4: 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  $f, g$     B. INFER  $f$  FROM  $g \circ f, g$

Figure 2. (A) Inductive and (B) deductive CCR.



# Compositional Causal Reasoning

## 4 Compositional Causal Reasoning

### Example 1: Decomposition of total causal effects in linear SCMs [17]

Let TE be the total effect, NDE the natural direct effect, and NIE the natural indirect effect. When causal functions are linear,

$$\underbrace{\text{TE}}_{\textit{global}} = \underbrace{\text{NDE}}_{\textit{local}} + \underbrace{\text{NIE}}_{\textit{local}}. \quad (3)$$



# Compositional Consistency Evaluation

## 5 Compositional Consistency Evaluation

---

Notation	Meaning
$\mathcal{M} := \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, p(\mathbf{u}) \rangle$	An SCM representing the problem.
$\mathcal{A}$	A model (e.g., an LM).
$\Phi$	The set of all causal measures.
$\varphi \in \Phi$	A measure of interest (e.g., the ATE).
$\varphi_{\mathbf{x}}$	A <i>causal query</i> about the value of $\varphi$ w.r.t. $\mathbf{X} \subset \mathbf{V}$ .
$\varphi_{\mathbf{x}}^*$	The true value of $\varphi_{\mathbf{x}}$ .

---



# Compositional Consistency Evaluation

## 5 Compositional Consistency Evaluation

### Definition 5: Compositional consistency

Inference is *compositionally consistent* when theoretically equivalent compositions are assessed to be equal.



# Compositional Consistency Evaluation

## 5 Compositional Consistency Evaluation

### Definition 6: External validity

Reasoning is *externally valid* when inferred quantities are equivalent to ground truth, up to some error  $\delta$ :

$$\theta(\varphi_{\mathbf{x}}^*, \widehat{\varphi}_{\mathbf{x}}) \leq \delta. \quad (4)$$

In Example 1, this entails that the following are less than  $\delta$ :

- $\theta(\text{TE}_{XY}^*, \widehat{\text{TE}}_{XY})$
- $\theta(\text{NDE}_{XY}^*, \widehat{\text{NDE}}_{XY})$
- $\theta(\text{TE}_{XY}^*, \widehat{\text{NDE}}_{XY} + \widehat{\text{NIE}}_{XY})$
- etc.



# Compositional Consistency Evaluation

## 5 Compositional Consistency Evaluation

### Definition 7: Internal consistency

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

$$\varphi_{\mathbf{x}}^* = \varphi_{\mathbf{x}'}^* \implies \theta(\widehat{\varphi}_{\mathbf{x}}, \widehat{\varphi}_{\mathbf{x}'}) \leq \delta. \quad (5)$$

Note that inferred quantities are compared to each other, not to ground truth. In Example 1, this entails that the following are less than  $\delta$ :

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



# Taxonomy of Reasoners

## 5 Compositional Consistency Evaluation

### Definition 8: Taxonomy of Reasoners

- 1. *Valid-consistent* (VC).
- 2. *Valid-inconsistent* (VI).
- 3. *Invalid-consistent* (IC).
- 4. *Invalid-inconsistent* (II).



# Compositional Consistency Evaluation

## 5 Compositional Consistency Evaluation

Each query is encoded as a **question template**

$$\mathcal{Q}_{\varphi_x} := (\varphi_x, \mathcal{S}), \quad (6)$$

where  $\varphi_x$  is implicit (i.e., not directly stated) and  $\mathcal{S}$  is the surface form that expresses accessory details (e.g., the background of a math word problem) [21].

$\mathcal{Q}_{\varphi_x}$  is expressed in a form comprehensible to  $\mathcal{A}$  (e.g., text, image, etc.).



# Compositional Consistency Evaluation

## 5 Compositional Consistency Evaluation

**Solutions** to causal queries are obtained by

$$\hat{\varphi}_{\mathbf{x}} := \mathcal{A}(\mathcal{Q}_{\varphi_{\mathbf{x}}}). \quad (7)$$

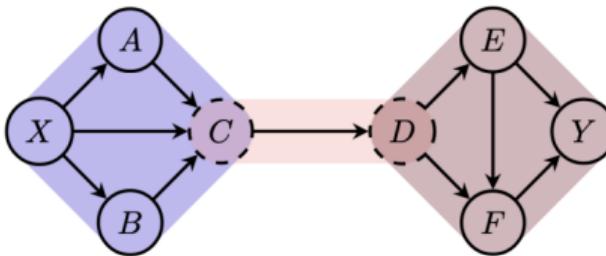
**Evaluation** entails computing approximation errors of form

$$\epsilon_{\varphi_{\mathbf{x}}} := \theta(\varphi_{\mathbf{x}}^*, \hat{\varphi}_{\mathbf{x}}) \quad (8)$$

or similar, for some metric  $\theta$ .

# Case Study: Graphs with Cutpoints

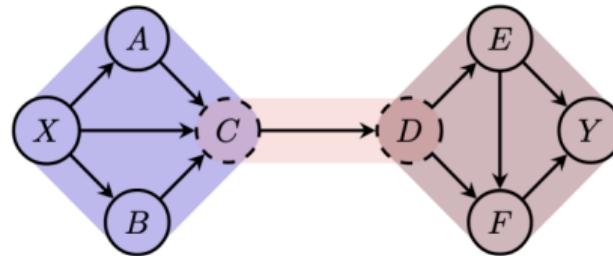
## 6 Inductive CCR Evaluation in Graphs with Cutpoints



- 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 [22].
  - E.g., the blue, pink, and maroon subgraphs.
- Removing a cutpoint disconnects the graph (e.g., nodes  $C, D$ ).

# Assumptions: Graphs with Cutpoints

## 6 Inductive CCR Evaluation in 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 BCCs

## 6 Inductive CCR Evaluation in Graphs with Cutpoints

### Theorem 1: PNS composition across 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

$$\text{PNS}_{XY} = \prod_{\{R_i, L_i\} \in \mathbf{C}} \text{PNS}_{R_i L_i} \quad (9)$$

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.



# Commutative Cut Trees

## 6 Inductive CCR Evaluation in Graphs with Cutpoints

### Definition 9: Commutative cut tree (CCT)

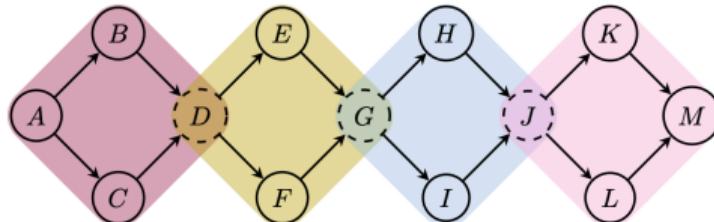
Let  $\mathcal{G}_{XY}$  be a causal graph satisfying A1–A4 and let  $\varphi$  be a causal measure that composes according to an associative function over BCCs (e.g., multiplication as in Theorem 1).

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

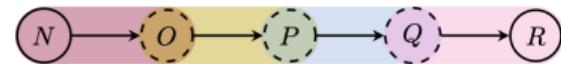
1. Construct a causal chain with nodes  $X \cup \mathbf{S} \cup Y$ , where  $\mathbf{S}$  is a topological ordering of the cutpoints in  $\mathcal{G}_{XY}$ .
2. Add a directed edge between any non-adjacent nodes in the chain to yield a complete graph where all directed paths point from root  $X$  to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

# Commutative Cut Trees

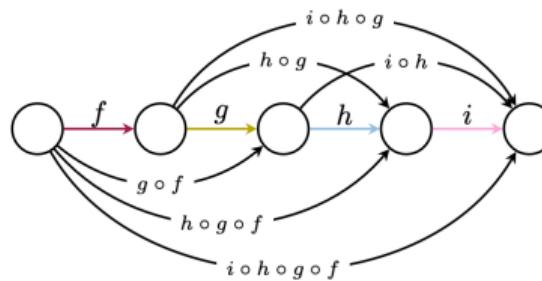
## 6 Inductive CCR Evaluation in Graphs with Cutpoints



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



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



C. CCT



# Commutative Cut Trees

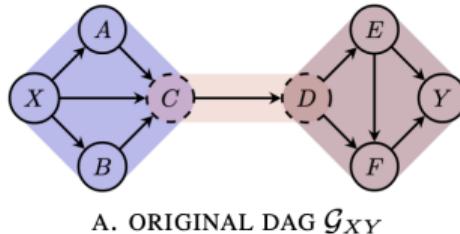
## 6 Inductive CCR Evaluation in Graphs with Cutpoints

CCTs as a useful abstraction:

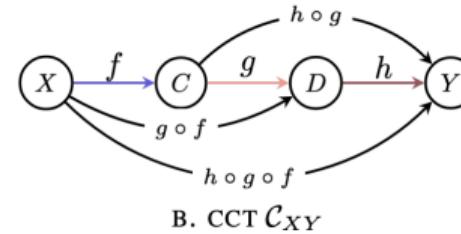
- Abstract away complexity in DAG by **collapsing BCCs** into single edges.
- Evaluate on complex DAGs with cutpoints as if they were **simply directed chains**.
- Simplify problem representation by (1) **marginalizing out variables** unnecessary for valid causal inference and (2) **visualizing pathways**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

# Running Example: Intuition for Algorithm 1

## 6 Inductive CCR Evaluation in Graphs with Cutpoints



A. ORIGINAL DAG  $\mathcal{G}_{XY}$



B. CCT  $\mathcal{C}_{XY}$

<i>Global</i>	$\text{PNS}_{XY}$
<i>Local</i>	$\text{PNS}_{XC}, \text{PNS}_{XD}, \text{PNS}_{CD}, \text{PNS}_{CY}, \text{PNS}_{DY}$
<i>Composition</i>	$\text{PNS}_{XC}\text{PNS}_{CY}, \text{PNS}_{XD}\text{PNS}_{DY}, \text{PNS}_{XC}\text{PNS}_{CD}\text{PNS}_{DY}$



# Models

## 7 Empirical Demonstration in LMs

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

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).



# Factual & Counterfactual Prompts

## 7 Empirical Demonstration 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.

- $\widehat{\text{PNS}}_{XC}$ : 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**: Llama 3.1, Llama 3.1 Math, and GPT-4o also queried with CoT formulation including one factual, one counterfactual example.



# Extracting and Evaluating PNS Values

## 7 Empirical Demonstration in LMs

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

$$RAE_{\text{external}} := \frac{|PNS^* - \widehat{PNS}|}{PNS^*}, \quad RAE_{\text{internal}} := \frac{|\widehat{PNS} - \widehat{PNS}'|}{\widehat{PNS}}. \quad (10)$$

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



## Remark

### 7 Empirical Demonstration in LMs

#### Remark 1

Success on CCR tasks is **necessary but not sufficient** for demonstrating that language models can reason.



# Results: Taxonomy of Reasoners

## 7 Empirical Demonstration in LMs

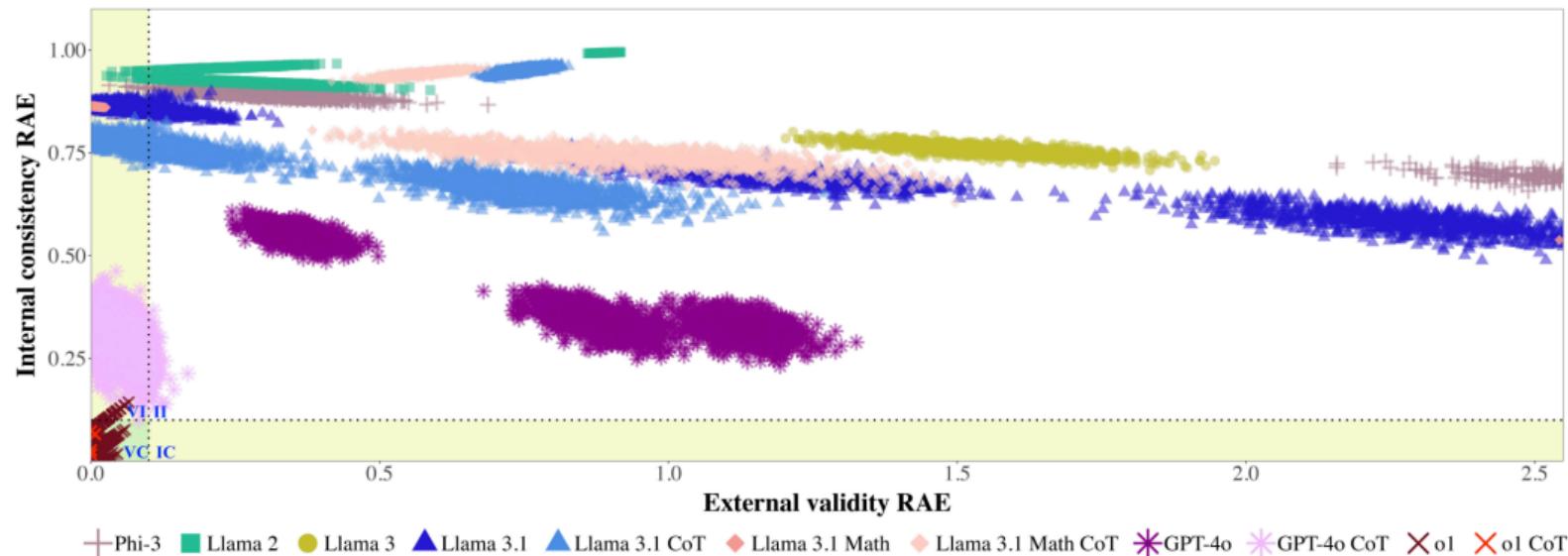
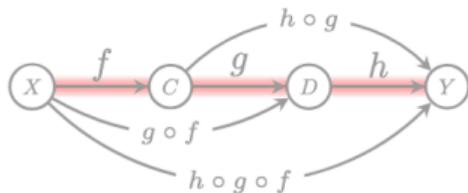


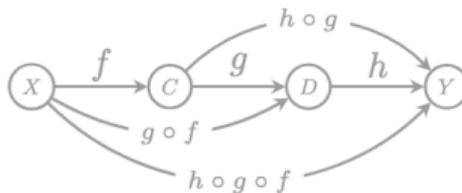
Figure 5. Composition RAE with respect to ground truth (external validity) and  $\widehat{PNS}_{XY}$  (internal consistency). Dotted lines represent the error threshold (RAE = 0.1), with reasoning quadrants VI/IC in yellow, VC in green, and II in white. Models are listed by increasing size (Table F.1). External validity RAE is truncated; for the full distribution, see Fig. F.6.

# Visualizing Reasoning with CCTs

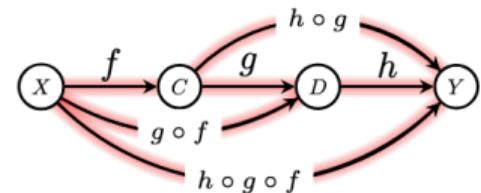
## 7 Empirical Demonstration in LMs



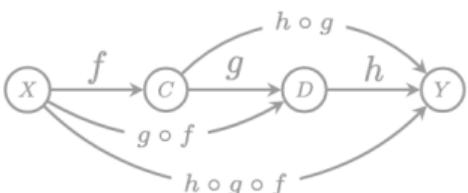
A. LLAMA 3.1 MATH



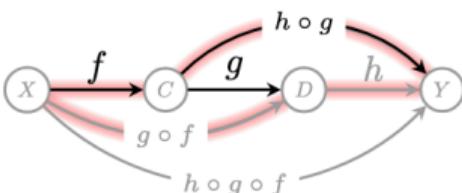
B. GPT-4O



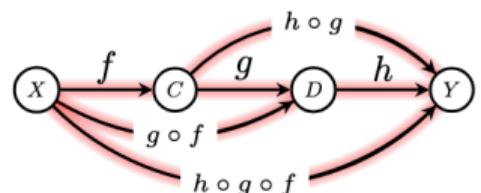
C. o1



D. LLAMA 3.1 MATH COT



E. GPT-4O COT

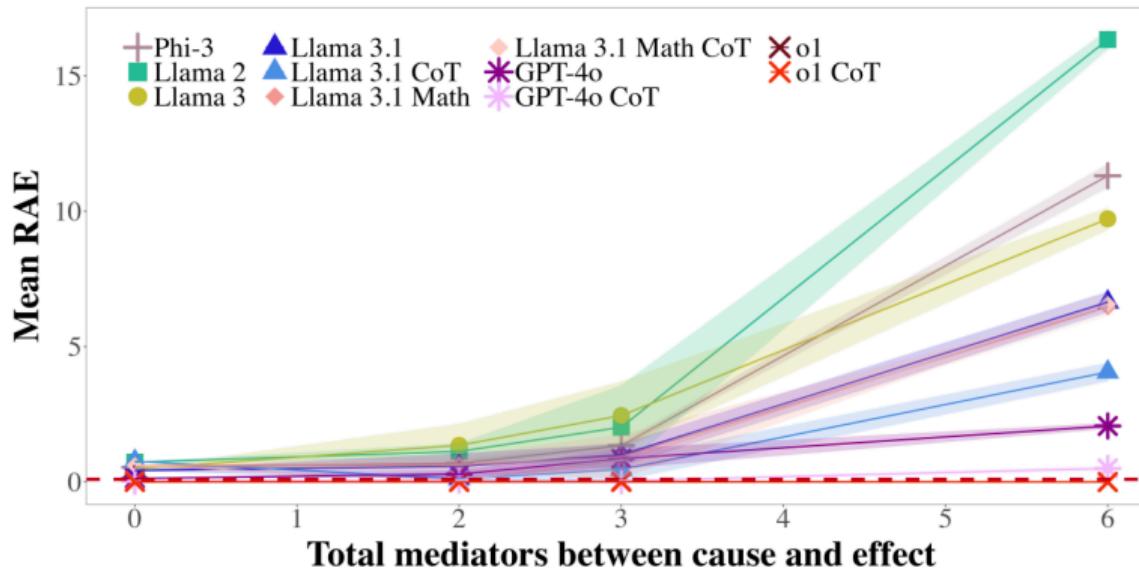


F. o1 COT



# Errors Increase With Mediation

## 7 Empirical Demonstration in LMs





# Limitations & Future Directions

## 8 Limitations & Future Directions

- Limited to one illustrative task as proof of viability.
- Future work could explore the automated design of tasks for large-scale benchmarking.
- Only considers the ATE and PNS under Theorem 1. Extensions could consider other estimands and compositional forms.



**Thank you! Any questions?**

[maasch@cs.cornell.edu](mailto:maasch@cs.cornell.edu)





## References

- [1] M. K. Goddu et al. "The development of human causal learning and reasoning". In: *Nature Reviews Psychology* (2024), pp. 1–21.
- [2] B. Schölkopf et al. "Toward causal representation learning". In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634.
- [3] J. Pearl. "Causality: Models, Reasoning, and Inference". In: Cambridge, UK: Cambridge University Press 19.2 (2000), p. 3.
- [4] B. M. Lake et al. "Building machines that learn and think like people". In: *Behavioral and brain sciences* 40 (2017), e253.
- [5] Z. Xu et al. "Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 25074–25087.
- [6] S. M. Frankland et al. "Concepts and compositionality: in search of the brain's language of thought". In: *Annual review of psychology* 71.1 (2020), pp. 273–303.
- [7] P. Schwartenbeck et al. "Generative replay underlies compositional inference in the hippocampal-prefrontal circuit". In: *Cell* 186.22 (2023), pp. 4885–4897.
- [8] B. M. Lake et al. "Human-like systematic generalization through a meta-learning neural network". In: *Nature* 623.7985 (2023), pp. 115–121.
- [9] J. Pearl. "Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach". In: *Proceedings, AAAI-82* (1982), pp. 133–136.



## References

- [10] G. R. Shafer et al. "Probability Propagation". In: *Annals of Mathematics and Artificial Intelligence* 2 (1990), pp. 327–351.
- [11] F. R. Kschischang et al. "Factor graphs and the sum-product algorithm". In: *IEEE Transactions on information theory* 47.2 (2001), pp. 498–519.
- [12] J. Pearl. "Interpretation and identification of causal mediation.". In: *Psychological methods* 19.4 (2014), p. 459.
- [13] T. J. VanderWeele. "Mediation analysis: a practitioner's guide". In: *Annual review of public health* 37 (2016), pp. 17–32.
- [14] 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](https://doi.org/10.1561/2200000106).
- [15] J. Pearl. "Causal diagrams for empirical research". In: *Biometrika* 82.4 (1995), pp. 669–688.
- [16] H. Jeong et al. "Finding and listing front-door adjustment sets". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 33173–33185.
- [17] J. Pearl. "Direct and Indirect Effects". In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco. 2001.
- [18] J. Pearl. "Probabilities of Causation: Three Counterfactual Interpretations and Their Identification". In: *Synthese* 121 (1999), pp. 93–149.
- [19] J. González et al. "Does Reasoning Emerge? Examining the Probabilities of Causation in Large Language Models". In: *Advances in Neural Information Processing Systems*. 2024.



## References

- [20] A. Hütük et al. "Reasoning Elicitation in Language Models via Counterfactual Feedback". In: *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. 2025.
- [21] A. Stolfo et al. "A Causal Framework to Quantify the Robustness of Mathematical Reasoning with Language Models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 545–561.
- [22] J. Westbrook et al. "Maintaining bridge-connected and biconnected components on-line". In: *Algorithmica* 7.1 (1992), pp. 433–464.



# Inductive CCR in Graphs with Cutpoints

---

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

---

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

**Output:** Reasoning errors  $\eta, \epsilon, \gamma$

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

*Compute quantity-wise errors.*

- 1: **for**  $\forall$  pairs  $\{R_i, L_{j>i}\}$  in  $\mathcal{C}_{XY}$  **do**
- 2:    $\eta_{R_i L_j} \leftarrow \theta(\varphi_{R_i L_j}^*, \hat{\varphi}_{R_i L_j})$                                $\triangleright$  External validity.

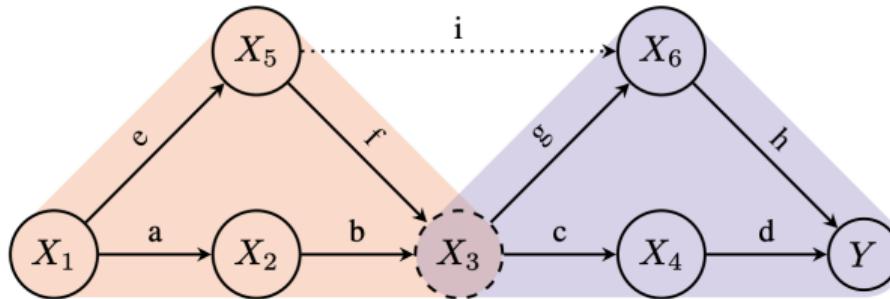
*Compute inductive reasoning errors.*

- 3: **for**  $\forall$  paths  $i$  from  $X$  to  $Y$  in  $\mathcal{C}_{XY}$  **do**
- 4:   Get composition  $\hat{\varphi}_i^\circ$  for path  $i$  from knowledge of edges  $j \in i$
- 5:    $\epsilon_i \leftarrow \theta(\varphi_{XY}^*, \hat{\varphi}_i^\circ)$                                        $\triangleright$  External validity.
- 6:    $\gamma_i \leftarrow \theta(\hat{\varphi}_{XY}, \hat{\varphi}_i^\circ)$      $\triangleright$  Internal consistency.

**return**  $\eta, \epsilon, \gamma$

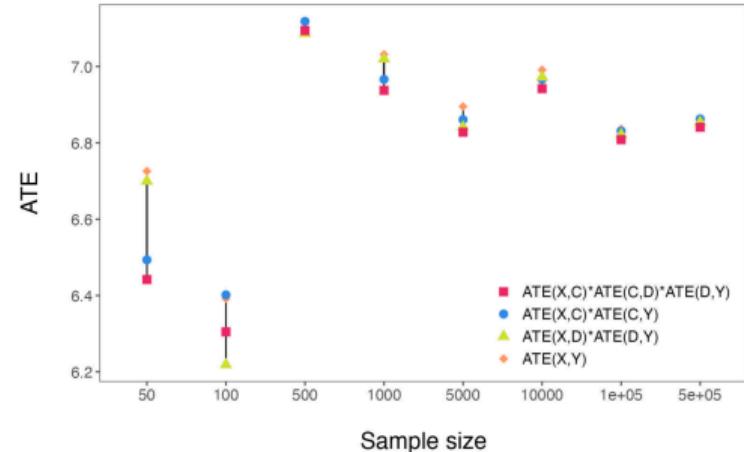
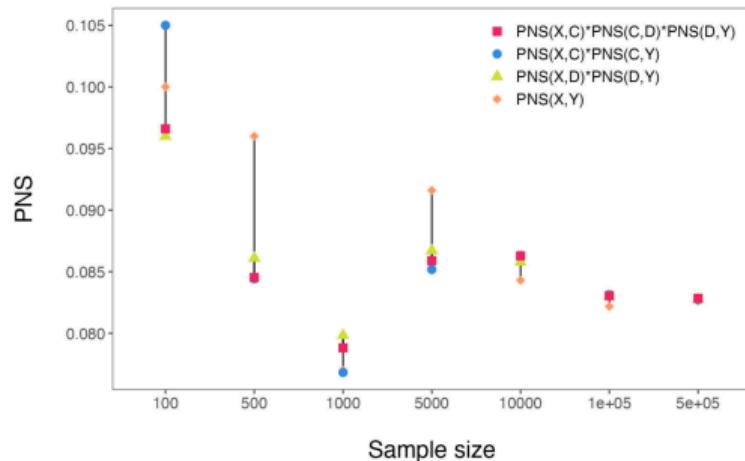
---

# 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  $X_5 \rightarrow X_6$  does not exist,  $\text{ATE}_{X_1 Y} = \text{ATE}_{X_1 X_3} \cdot \text{ATE}_{X_3 Y}$ .
- If  $X_5 \rightarrow X_6$  does exist, then product is summed with additional term corresponding to the path-specific effect for  $X_1 \rightarrow X_5 \rightarrow X_6 \rightarrow Y$ , which does not pass through  $X_3$ .

# PNS & ATE Composition Across BCCs





# 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.*