



Compositional Causal Reasoning Evaluation in Language Models

Presenter: Jacqueline Maasch

maasch@cs.cornell.edu | [arXiv:2503.04556](https://arxiv.org/abs/2503.04556)

Jacqueline Maasch¹, Alihan Hüyük², Xinnuo Xu³, Aditya Nori³, Javier González³

¹Cornell Tech, ²Harvard University, ³Microsoft Research Cambridge



Presentation Overview

1 Background

1. Background.
2. Conceptual Framework:
 - Compositional Causal Reasoning (CCR).
 - Compositional Consistency Evaluation.
3. Algorithm: Inductive CCR Evaluation in Graphs with Cutpoints.
4. Results: Empirical Demonstration in Language Models.



Background



Reasoning in AI

1 Background

- What is reasoning?
 - **Process** of applying logic to draw valid conclusions from new or prior information, with the aim of seeking *truth* (Wikipedia).
 - **Activity** of forming new beliefs based on existing beliefs and new evidence.
 - Bayesian inference: $P(H | E) = \frac{P(E|H)P(H)}{P(E)}$, where H is hypothesis and E is evidence.
 - Etc.
- **Many forms**: quantitative, logical, visual, spatial, moral, legal, etc.
- We consider two forms in tandem: **causal reasoning** and **compositional reasoning**.



Reasoning in AI: Two Issues

1 Background

Increase reasoning capacities:

Published as a conference paper at ICLR 2025

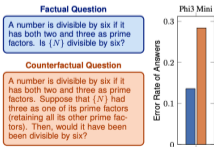
REASONING ELICITATION IN LANGUAGE MODELS VIA COUNTERFACTUAL FEEDBACK

Alihan Hüyük,^{*1} Xinnuo Xu,[‡] Jacqueline Maasch,[§] Aditya V. Nori,[‡] Javier González[‡]

¹Harvard 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



Principled evaluation frameworks:

Compositional Causal Reasoning Evaluation in Language Models

Jacqueline R. M. A. Maasch¹ Alihan Hüyük² Xinnuo Xu³ Aditya V. Nori³ Javier Gonzalez³

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.

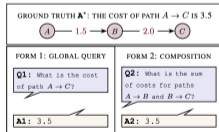


Figure 1. Compositionally consistent responses to two formulations of a simple (non-causal) query. Reasoning is externally valid if $A1$ and $A2$ both equal A^* , and internally consistent if $A1=A2$.

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.



Causal Reasoning

1 Background

- Causal reasoning is a defining outcome of human evolution [1].
- Enables humans and machines 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 **copng with complexity**.

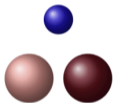


Compositionality is Ubiquitous

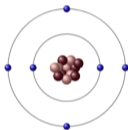
1 Background

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

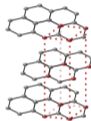
Physics.



Subatomic particles.

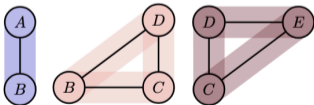


Carbon-12 atom.

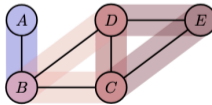


Graphite structure.

Graphical modeling.



Local subgraphs.



Global graph.



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

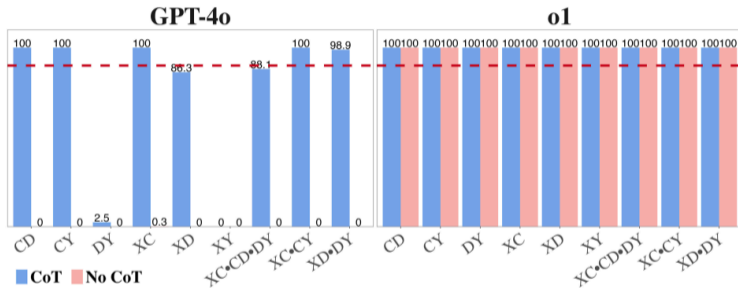
2 Our Work

1. **A compositional view of causal reasoning in LMs.** The ability 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



- A simple CCR task revealed **taxonomically distinct error patterns**.
- Only o1 was fully correct on this task.



Framework: Compositional Causal Reasoning Evaluation



Preliminaries: Causal Models

3 Framework: Compositional Causal Reasoning Evaluation

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

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

- $\mathbf{U} = \{U_i\}_{i=1}^n$ 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 Framework: Compositional Causal Reasoning Evaluation

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 Framework: Compositional Causal Reasoning Evaluation

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 Framework: Compositional Causal Reasoning Evaluation

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 Framework: Compositional Causal Reasoning Evaluation

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.



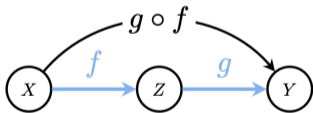
Compositional Causal Reasoning

3 Framework: Compositional Causal Reasoning Evaluation

Definition 4: Compositional Causal Reasoning (CCR)

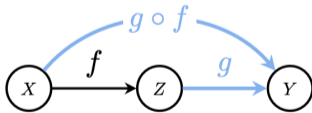
The ability to infer compositions and decompositions of causal measures in factual and counterfactual worlds.

Inductive CCR:



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

Deductive CCR:



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



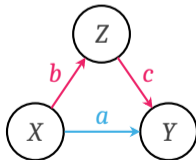
A Classic Example

3 Framework: Compositional Causal Reasoning Evaluation

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,

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



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



Compositional Consistency Evaluation

3 Framework: Compositional Causal Reasoning Evaluation

Definition 5: Compositional consistency

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



Compositional Consistency Evaluation

3 Framework: Compositional Causal Reasoning 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).
Φ	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 φ w.r.t. $\mathbf{X} \subset \mathbf{V}$.
$\varphi_{\mathbf{x}}^*$	The true value of $\varphi_{\mathbf{x}}$.



Compositional Consistency Evaluation

3 Framework: Compositional Causal Reasoning Evaluation

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

$$Q_{\varphi_{\mathbf{x}}} := (\varphi_{\mathbf{x}}, \mathcal{S}), \quad (4)$$

where $\varphi_{\mathbf{x}}$ 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].

$Q_{\varphi_{\mathbf{x}}}$ is expressed in a form comprehensible to \mathcal{A} (e.g., text, image, etc.).



Compositional Consistency Evaluation

3 Framework: Compositional Causal Reasoning Evaluation

Solutions to causal queries are obtained by

$$\hat{\varphi}_{\mathbf{x}} := \mathcal{A}(\mathcal{Q}_{\varphi_{\mathbf{x}}}). \quad (5)$$

Evaluation entails computing approximation errors of form

$$\epsilon_{\varphi_{\mathbf{x}}} := \theta(\varphi_{\mathbf{x}}^*, \hat{\varphi}_{\mathbf{x}}) \quad (6)$$

or similar, for some metric θ .



Compositional Consistency Evaluation

3 Framework: Compositional Causal Reasoning Evaluation

Definition 6: External validity

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

$$\theta(\varphi_{\mathbf{x}}^*, \widehat{\varphi}_{\mathbf{x}}) \leq \delta. \quad (7)$$

In Example 1, this entails that the following are less than δ :

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



Compositional Consistency Evaluation

3 Framework: Compositional Causal Reasoning Evaluation

Definition 7: Internal consistency

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

$$\varphi_{\mathbf{x}}^* = \varphi_{\mathbf{x}'}^* \Rightarrow \theta(\widehat{\varphi}_{\mathbf{x}}, \widehat{\varphi}_{\mathbf{x}'}) \leq \delta. \quad (8)$$

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

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



Taxonomy of Reasoners

3 Framework: Compositional Causal Reasoning Evaluation

Definition 8: Taxonomy of Reasoners

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

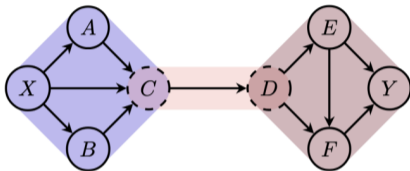


Algorithm: Inductive CCR Evaluation in Graphs with Cutpoints



Case Study: Graphs with Cutpoints

4 Algorithm: Inductive CCR Evaluation in Graphs with Cutpoints

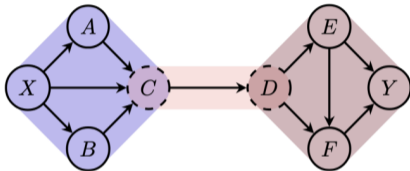


- A **cutpoint** is any node contained in multiple **biconnected components** (BCCs):
 - Maximal biconnected subgraphs induced by a partition of edges. Two edges are in the same partition if and only if they share a common simple cycle [22].
 - E.g., the blue, pink, and maroon subgraphs.
- Removing a cutpoint disconnects the graph (e.g., nodes C, D).



Assumptions: Graphs with Cutpoints

4 Algorithm: 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

4 Algorithm: 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

4 Algorithm: 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 φ be a causal measure that composes according to an associative function over BCCs (e.g., multiplication as in Theorem 1).

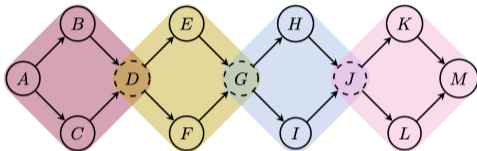
CCT \mathcal{C}_{XY} is a transformation of \mathcal{G}_{XY} that models all CCR pathways from root X to leaf Y for measure φ . \mathcal{C}_{XY} is obtained by a two-step transformation of \mathcal{G}_{XY} :

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 .

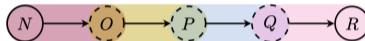


Commutative Cut Trees

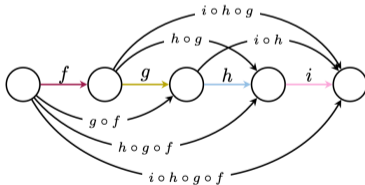
4 Algorithm: Inductive CCR Evaluation in Graphs with Cutpoints



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



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



C. CCT



CCTs: A Useful Abstraction

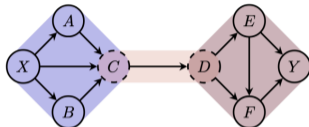
4 Algorithm: Inductive CCR Evaluation in Graphs with Cutpoints

- 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**.

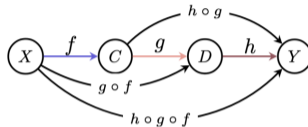


Running Example: Intuition for Algorithm 1

4 Algorithm: Inductive CCR Evaluation in Graphs with Cutpoints



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



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

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



Results: Empirical Demonstration in LMs



Models

5 Results: Empirical Demonstration in LMs

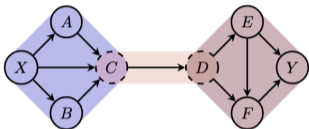
MODEL	PARAMETERS	LINK
Phi-3-Mini-128K-Instruct (Abdin et al., 2024)	3.82B	https://huggingface.co/microsoft/Phi-3-mini-128k-instruct
Llama-2-7b-Chat-HF (Touvron et al., 2023)	6.74B	https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
Llama-3-8B-Instruct (Dubey et al., 2024)	8.03B	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
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).

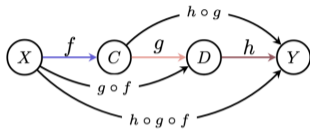


Factual & Counterfactual Prompts

5 Results: Empirical Demonstration in LMs



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



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

- SCM $\mathcal{M} := \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, p(\mathbf{u}) \rangle$: \mathbf{V} are binary, $f \in \mathcal{F}$ are logical *or* (\vee), $p(\mathbf{u})$ is Bernoulli.
- Logical *or* is a monotone boolean function.
- Each node in the DAG is a person in our word problem: $X = \text{Xinyu}$, $A = \text{Ara}$, $B = \text{Becca}$, $C = \text{Celine}$, $D = \text{Daphne}$, $E = \text{Emma}$, $F = \text{Fox}$, $Y = \text{Yasmin}$.

$$v_i = pa_1, \vee \dots \vee pa_k \vee \text{Ber}(0.7) \quad (10)$$



Factual & Counterfactual Prompts

5 Results: 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{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{PNS}_{DY} : Interventions on D (Daphne's happiness), queries on Y (Yasmin's happiness).
- **CoT formulation**: Demonstrated one factual and one counterfactual example.



Extracting and Evaluating PNS Values

5 Results: Empirical Demonstration in LMs

- 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).

$$RAE_{\text{external}} := \frac{|PNS^{* \cdot} - \widehat{PNS}^{\cdot}|}{PNS^{* \cdot}}, \quad RAE_{\text{internal}} := \frac{|\widehat{PNS}^{\cdot} - \widehat{PNS}^{\cdot'}|}{\widehat{PNS}^{\cdot}}. \quad (11)$$

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



Remark

5 Results: 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

5 Results: Empirical Demonstration in LMs

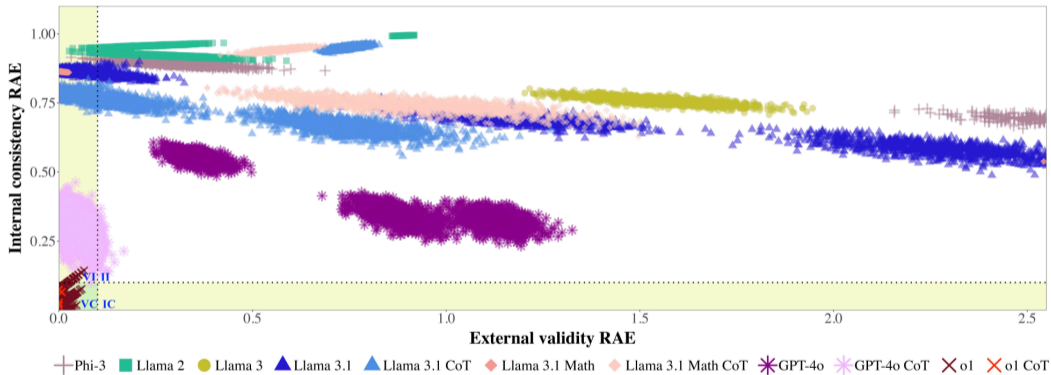
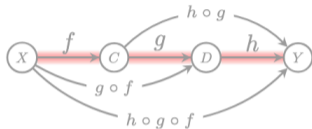


Figure 5. Composition RAE with respect to ground truth (external validity) and $\widehat{\text{PNS}}_{XY}$ (internal consistency). Dotted lines represent the error threshold (RAE = 0.1), with reasoning quadrants VI/IC in yellow, VC in green, and II in white. Models are listed by increasing size (Table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

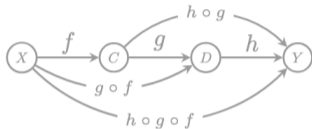


Visualizing Reasoning with CCTs

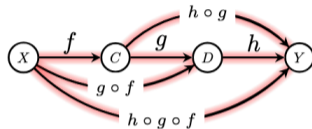
5 Results: Empirical Demonstration in LMs



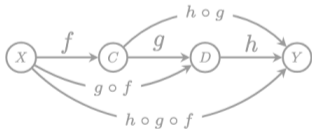
A. LLAMA 3.1 MATH



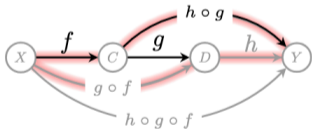
B. GPT-4O



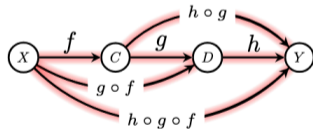
C. o1



D. LLAMA 3.1 MATH COT



E. GPT-4O COT

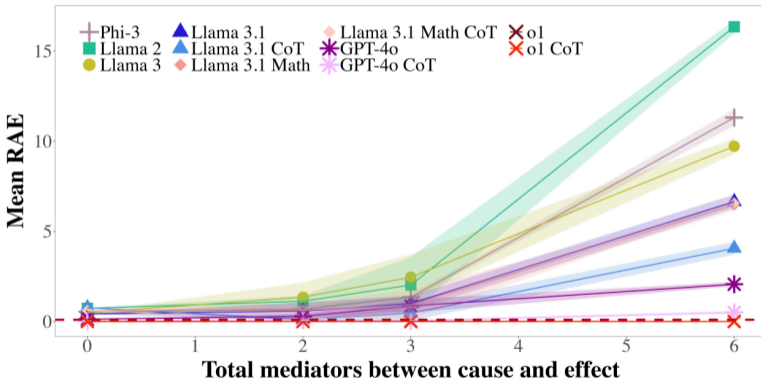


F. o1 COT



Errors Increase With Mediation

5 Results: Empirical Demonstration in LMs





Limitations & Future Directions

6 Limitations & Future Directions

- Limited to one illustrative task as proof of viability.
- Future work: automated task design 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`





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.



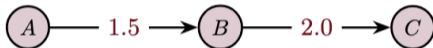
References

- [20] A. Hüyük 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.



A Noncausal Example for Intuition

GROUND TRUTH \mathbf{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



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.

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

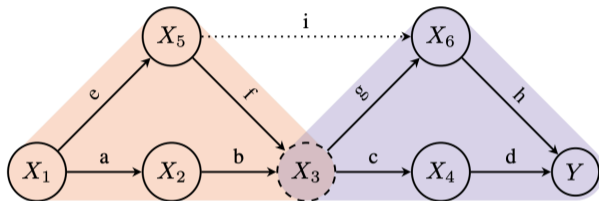
Compute inductive reasoning errors.

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

return η, ϵ, γ



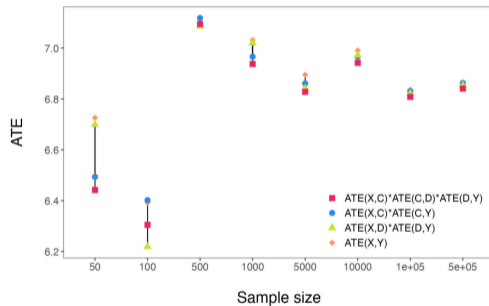
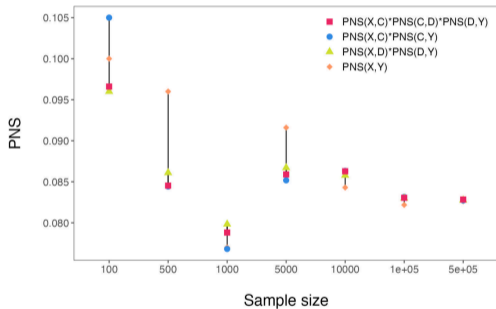
ATE Composition Across BCCs



- 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, $ATE_{X_1 Y} = ATE_{X_1 X_3} \cdot ATE_{X_3 Y}$.
- 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.