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.

1. Introduction

Causal reasoning is a defining outcome of human evolution (Goddu & Gopnik, 2024). Humans flexibly reason about cause and effect in factual realities that can be observed and intervened on, as well as imagined counterfactual worlds. A causal lens enables humans and machines alike to learn generalizable lessons about the mechanics of the universe (Schölkopf et al., 2021). Thus, human-like AI might require reasoning at all three levels of Pearl's Causal Hierarchy: associational, interventional, and counterfactual (Pearl, 2000).

Human-like AI might also require *compositional reasoning* (Lake et al., 2017): the capacity to recognize and synthesize novel combinations of previously observed concepts (Xu et al., 2022). Compositionality is ubiquitous in the physical world, symbolic systems, and human cognition (Frankland & Greene, 2020), underlying both visual perception (Schwartenbeck et al., 2023) and language (Lake & 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.

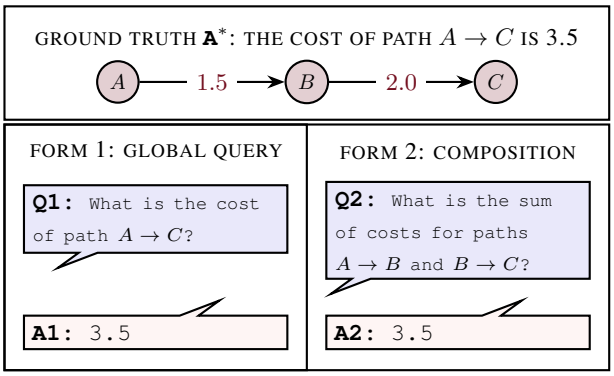


Figure 1. Compositionally consistent responses to two formulations of a simple (non-causal) query. Reasoning is externally valid if A1 and A2 both equal A^*, and internally consistent if A1==A2.

The present work explores causal and compositional reasoning in tandem. We center our focus on reasoning evaluation in language models (LMs), given increasing interest in LM reasoning emergence (Huang & Chang, 2023; Qiao et al., 2023; Mialon et al., 2023). Following from traditions in causal inference and graphical modeling, we define *compositional causal reasoning* (CCR) as

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

By extension, this requires reasoning over the propagation of causal quantities through graphs. To facilitate CCR evaluation, we introduce a framework for the exhaustive assessment of *compositional consistency*: correct inference that equivalent compositions are indeed equal. We measure compositional consistency with respect to ground truth (*external validity*) and concordance among the LM's responses (*internal consistency*) (Fig. 1). We empirically demonstrate instantiations of our framework for two causal measures: the average treatment effect (ATE) and the probability of necessity and sufficiency (PNS; Pearl 1999), which coincides with the ATE in certain data generating processes.

[1]Cornell Tech [2]Harvard University [3]Microsoft Research Cambridge. Correspondence to: Jacqueline Maasch <maasch@cs.cornell.edu>.

[1]See Fig. A.1 for examples.

1.1. Contributions

- §3 **A compositional view of causal reasoning in LMs.** We formally express CCR as the ability of an LM to infer causal measure compositions (*inductive* reasoning) and decompositions (*deductive* reasoning).
- §4 **Metrics and reasoning taxonomy.** Following from our proposed measures of external validity and internal consistency, we introduce four categories of reasoners: *valid-consistent* (VC), *valid-inconsistent* (VI), *invalid-consistent* (IC), and *invalid-inconsistent* (II).
- §5 **An evaluation framework.** We introduce a procedure for evaluating inductive CCR for the ATE and PNS in causal graphs with cutpoints (Alg. 1).
- §6 **Preliminary empirical demonstration.**[2] We deploy Alg. 1 to evaluate CCR in seven LM architectures, with and without chain-of-thought (CoT) prompting. Even on a simple CCR problem, our framework revealed taxonomically distinct patterns of inconsistent and invalid reasoning ranging from II to VC.

1.2. Related Works

LM Reasoning This work contributes to the theoretical and empirical study of compositional (Hudson & Manning, 2019; Hupkes et al., 2020; Xu et al., 2022; Ito et al., 2022; Hsieh et al., 2023), causal (Wang et al., 2023c; Du et al., 2022), mathematical (Saxton et al., 2019; Lewkowycz et al., 2022; Stolfo et al., 2023; Lu et al., 2023b), inductive (Qiu et al., 2024), and graphical reasoning (Wang et al., 2023a; He et al., 2024) in AI. Compositional reasoning has been examined in LMs (Li et al., 2023; Lu et al., 2023a; Dziri et al., 2023), multimodal LMs (Wu et al., 2024), and diffusion models (Du et al., 2023; Okawa et al., 2023; Su et al., 2024). In this work, we focus on autoregressive LMs. Consistency is an ongoing concern in LM research (Wang et al., 2023b; Li et al., 2024), and this work introduces a new compositional framework for its measurement.

Despite some optimistic results (Kıcıman et al., 2023), several recent works hypothesize that current LMs are not capable of true logical reasoning (Mirzadeh et al., 2024) and are merely *causal parrots* (Zečević et al., 2023). Counterfactual reasoning in LMs can be brittle (González & Nori, 2024), and performance on formal causal reasoning tasks can decline monotonically with task difficulty (Jin et al., 2023). Similarly, multiple works find that models struggle with compositional reasoning in vision and language, especially for complex compositions (Agrawal et al., 2017; Ma et al., 2023; Ray et al., 2023; Press et al., 2023). Efforts to elicit causal, compositional, and mathematical reasoning via fine-tuning (Hüyük et al., 2025) or CoT prompting (Wei et al., 2022; Jin et al., 2023; Press et al., 2023) have shown promising yet limited success. A survey of causal reasoning benchmarks found that most suffer from critical design flaws (e.g., data contamination and inappropriate evaluation metrics), highlighting the need for improved evaluation frameworks (Yang et al., 2024). To our knowledge, prior frameworks do not explicitly and systematically incorporate causal compositionality. The present work lays a foundation for probing the compositional consistency of causal reasoning, with preliminary results comparing CoT and non-CoT prompting. See Appendix A for additional related works.

Compositionality & Causality While explicitly combining compositional and causal reasoning evaluation in LMs is new, the intersection of compositionality and causality already enjoys a rich mathematical framework: the formalisms and methods of graphical modeling and causal inference. A central concern of probabilistic and causal graphical modeling is factorization: the expression of complex multivariate distributions as products (or *compositions*) of local distributions, enabling efficient learning and inference (Koller & Friedman, 2009). Graphs provide expressive representations for joint distributions, their factors, and the propagation of quantities through systems (Pearl, 1982; Shafer & Shenoy, 1990; Kschischang et al., 2001). In causal inference, the decomposition of causal effects (Avin et al., 2005; Pearl, 2014; VanderWeele, 2014; Singal & Michailidis, 2024) plays a central role in mediation analysis (VanderWeele, 2016), fairness analysis (Plečko & Bareinboim, 2024), and covariate adjustment in the presence of latent variables (Pearl, 1995; Jeong et al., 2022). These traditions offer a convenient mathematical language for evaluating compositional and causal reasoning simultaneously.

2. Preliminaries

Notation Uppercase denotes univariate random variables (e.g., Y) and bold uppercase denotes sets or multivariate random variables (e.g., $\mathbf{V}$), with realizations in lowercase (e.g., scalars x; vector values $\mathbf{x}$). Models and graphs are denoted by calligraphic script (e.g., causal graph $\mathcal{G}$). Probability measures are denoted by $\mathbb{P}$ and probability distributions by p. For $X \in \mathcal{G}$, we denote the parents of X by $\mathbf{pa}_X$.

2.1. Causal Models

Structural Causal Models Structural causal models (SCMs; Pearl 2009) provide a convenient coupling for our framework: (1) a rich mathematical language for expressing compositionality in terms of causal measure compositions and (2) an intuitive visual language in the form of directed acyclic graphs (DAGs).

Definition 2.1 (Structural causal model (SCM), Pearl 2001)**.** An SCM is a tuple $\mathcal{M} := \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, p(\mathbf{u}) \rangle$, where $\mathbf{U} = \{U_i\}_{i=1}^n$ is a set of exogenous variables determined by factors outside $\mathcal{M}$, $\mathbf{V} = \{V_i\}_{i=1}^n$ is a set of observed endogenous variables determined by variables in $\mathbf{U} \cup \mathbf{V}$,

[2] Data and code will be publicly released upon publication.

$\mathcal{F} = \{f_i\}_{i=1}^n$ is a set of structural functions such that $V_i = f_i(\mathbf{pa}_{V_i}, U_i)$, and $p(\mathbf{u})$ is the distribution over $\mathbf{U}$.

We restrict our attention to *positive-Markovian* SCMs whose graphical representations are DAGs.

Definition 2.2 (Positive-Markovian SCM, Pearl 1999)**.** An SCM is Markovian if its graphical representation is acyclic and exogenous variables U_i are mutually independent. An SCM is positive-Markovian if it is Markovian and $\mathbb{P}(v) > 0$ for every realization $V \in \mathbf{V} = v$.

Let $X, Y \in \mathcal{M}$ be binary random variables. Let pa_X denote the realization of a parent of X. Under Def. 2.2, the causal effect of X on Y is identifiable as (Pearl, 1999)

$$\mathbb{P}(Y = y \mid do(X = x)) = \sum_{pa_X} \mathbb{P}(y \mid x, pa_X)\mathbb{P}(pa_X). \quad (1)$$

2.2. Causal Measures

Average Treatment Effect The most widely studied causal estimand (Imbens, 2004), the ATE measures the effect of receiving treatment versus no treatment on the mean outcome over the population.

Definition 2.3 (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:

$$\text{ATE} \coloneqq \mathbb{E}[Y \mid do(X = 1)] - \mathbb{E}[Y \mid do(X = 0)]. \quad (2)$$

Probability of Necessity and Sufficiency In propositional logic, we say that X is *necessary* for Y when $Y \implies X$, X is *sufficient* for Y when $X \implies Y$, and X is *necessary and sufficient* for Y when $X \iff Y$. Pearl (1999) introduced a probabilistic framework for reasoning over necessity and sufficiency with the *probabilities of causation* (PrC): the probabilities of necessity (PN), sufficiency (PS), and necessity and sufficiency (PNS). In the present work, we focus on the PNS.

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 $Y = \text{FALSE}$.

Definition 2.4 (Probability of necessity and sufficiency (PNS), Pearl 1999)**.** The probability that x is necessary and sufficient to produce y is given as

$$\text{PNS} \coloneqq \mathbb{P}(y_x, y'_{x'}) = \mathbb{P}(x, y)PN + \mathbb{P}(x', y')PS. \quad (3)$$

The PNS is point identifiable from causal effects when Y is monotonic in X: changing X from FALSE to TRUE does not induce Y to change from TRUE to FALSE.

Definition 2.5 (Identification of the PNS, Tian & Pearl 2000)**.** Given a positive-Markovian SCM (Def. 2.2) for which Y is monotonic in X, the PNS is given as

$$\mathbb{P}(y_x) - \mathbb{P}(y_{x'}) = \mathbb{P}(y \mid do(x)) - \mathbb{P}(y \mid do(x')) \quad (4)$$

where effects $\mathbb{P}(y_x)$ and $\mathbb{P}(y_{x'})$ are identifiable by Eq. 1. Note that Eq. 4 is equivalent to the ATE (Proposition B.4). We will leverage this fact for CCR evaluation in Section 5.

3. Compositional Causal Reasoning

Compositionality has been variously defined in linguistics (Haugeland, 1979; Wittgenstein, 2009), category theory (Fong & Spivak, 2019), quantum theory (Coecke, 2023), and other domains. For example, a Schrödinger compositional theory dictates that compositions are "greater than the sum of their parts," in some sense (Coecke, 2023). However, such emergent effects are not relevant in our setting. We highlight this definitional variation to clarify that measuring compositional reasoning in AI is contingent on a chosen definition of compositionality. We select a function-based viewpoint that draws from probabilistic graphical modeling and causal inference.[3]

Definition 3.1. *Compositionality* exists when a measure f can be expressed as a function of measures $\{g_i\}_{i=1}^{n\geq 2}$.

This definition is intentionally lax to capture a wide range of mathematical behavior. It allows for any function to serve as the compositional rule (e.g., addition, multiplication, etc.). As in probabilistic graphical modeling, we refer to compositions f as *global* measures and to constituent g_i as *local* measures. These are terms of relative scale: a local measure in one setting may also admit decompositions at finer granularity. Thus, compositionality implies a scale of interest (as exemplified by the physics example in Fig. A.1). Def. 3.1 arises frequently in causal inference, e.g.:

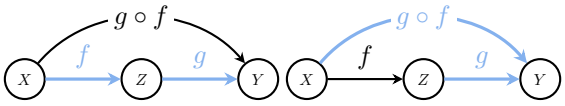
Example 3.2 (Decomposition of total causal effects in linear SCMs, Pearl 2001). Let TE be the total effect, NDE the natural direct effect, and NIE the natural indirect effect. When causal functions are linear,

$$\underbrace{\text{TE}}_{\textit{global}} = \underbrace{\text{NDE}}_{\textit{local}} + \underbrace{\text{NIE}}_{\textit{local}}. \quad (5)$$

Following from Definition 3.1, we define a compositional interpretation of causal reasoning in AI.

Definition 3.3 (Compositional causal reasoning (CCR))**.** The ability to correctly infer (1) how local causal measures *compose* into global causal measures and (2) how global causal measures *decompose* into local causal measures, in both factual and counterfactual worlds.

[3] Our definition is similar to *decomposability* as given by Def. 3.5 in Plečko & Bareinboim (2024).



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

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

Definition 3.3 encompasses reasoning over both compositions and decompositions, which we can disaggregate into *inductive* and *deductive* reasoning (Fig. 2).

Definition 3.4 (Inductive CCR)**.** The ability to reason about the *composition* of causal measures. With respect to Def. 3.1, this corresponds to inferring composition f given knowledge of local measures $\{g_i\}_{i=1}^{n\geq 2}$.

Definition 3.5 (Deductive CCR)**.** The ability to reason about the *decomposition* of causal measures. With respect to Def. 3.1, this corresponds to inferring local measure g_j given knowledge of composition f and measures $\{g_i\}_{i=1}^{n\geq 2} \setminus g_j$.

With respect to Example 3.2, inductive CCR entails inferring TE from NDE and NIE, while deductive CCR entails inferring NDE from TE and NIE (etc.).

4. Compositional Consistency Evaluation

4.1. Evaluation by Task-Metric-Model

We tether evaluation to a *task-metric-model triple* (Schaeffer et al., 2023). Let $\mathcal{A}$ be a model (e.g., an LM). Let Φ be the set of all causal measures. Let $\varphi \in \Phi$ be a measure of interest that we seek to evaluate for variables $\mathbf{V}$ in SCM $\mathcal{M}$ (e.g., the ATE). Let $\varphi_{\mathbf{x}}$ be a *causal query* about the value of φ with respect to (w.r.t.) $\mathbf{X} \subset \mathbf{V}$. We denote the true value of $\varphi_{\mathbf{x}}$ by $\varphi^*_{\mathbf{x}}$. Each query is encoded as a *question template*

$$\mathcal{Q}_{\varphi_{\mathbf{x}}} := (\varphi_{\mathbf{x}}, \mathcal{S}), \tag{6}$$

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) (Stolfo et al., 2023). $\mathcal{Q}_{\varphi_{\mathbf{x}}}$ is expressed in a form comprehensible to $\mathcal{A}$ (e.g., text, image, etc.). Solutions to causal queries are obtained by

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

Evaluation entails computing approximation errors of form $\epsilon_{\varphi_{\mathbf{x}}} := \theta(\varphi^*_{\mathbf{x}}, \widehat{\varphi}_{\mathbf{x}})$ or similar, for some metric θ.

CCR evaluation requires $\mathcal{A}$ to answer queries for a set of cause-effect pairs. Let $\mathcal{Q} := \{\mathcal{Q}_{\varphi_{\mathbf{x}}}\} \forall \varphi_{\mathbf{x}}$ of interest. We define a CCR task as the tuple $\mathcal{T} := \langle \varphi, \mathcal{M}, \mathcal{Q} \rangle$ and a task-metric-model triple as $\langle \mathcal{T}, \theta, \mathcal{A} \rangle$. Multiple φ may be identifiable for $\mathcal{M}$, and infinitely many $\mathcal{Q}$ can map to $\{\varphi, \mathcal{M}\}$. Further, success on $\langle \varphi, \mathcal{M}, \mathcal{Q} \rangle$ does not imply success on $\langle \varphi, \mathcal{M}, \mathcal{Q}' \rangle$, $\langle \varphi, \mathcal{M}', \mathcal{Q}' \rangle$, $\langle \varphi', \mathcal{M}, \mathcal{Q}' \rangle$, etc.

4.2. Metrics for Compositional Consistency

We now explore a prime benefit of compositional perspectives on reasoning: the ease of introducing notions of *consistency*. The following concept of consistency can be applied to any form of reasoning, not only causal.

Definition 4.1 (Compositional consistency)**.** Inference is *compositionally consistent* when theoretically equivalent compositions are assessed to be equal.

Under the umbrella of compositional consistency, we can quantify CCR in many ways. For example, a model that succeeds at task $\langle \varphi, \mathcal{M}, \mathcal{Q} \rangle$ should not be prone to false negatives (i.e., failing to infer compositionality when it exists) nor false positives (i.e., hallucinating compositionality when it does not exist). In this work, we emphasize the *external validity* and *internal consistency* of CCR.

Definition 4.2 (External validity)**.** Reasoning is *externally valid* when inferred quantities are equivalent to ground truth, up to some error δ:

$$\theta(\varphi^*_{\mathbf{x}}, \widehat{\varphi}_{\mathbf{x}}) \leq \delta. \tag{8}$$

In Example 3.2, externally valid reasoning for cause-effect pair $\{X, Y\}$ entails that the following are below some threshold: $\theta(\text{TE}^*_{XY}, \widehat{\text{TE}}_{XY})$, $\theta(\text{NDE}^*_{XY}, \widehat{NDE}_{XY})$, $\theta(\text{TE}^*_{XY}, \widehat{\text{NDE}}_{XY} + \widehat{\text{NIE}}_{XY})$, etc.

Definition 4.3 (Internal consistency)**.** Reasoning is *internally consistent* when quantities that are theoretically equivalent are inferred to be equivalent, up to some error δ:

$$\varphi^*_{\mathbf{x}} = \varphi^*_{\mathbf{x}'} \implies \theta(\widehat{\varphi}_{\mathbf{x}}, \widehat{\varphi}_{\mathbf{x}'}) \leq \delta. \tag{9}$$

Note that inferred quantities are compared to each other, not to ground truth. In Example 3.2, internally consistent reasoning entails that $\theta(\widehat{\text{TE}}_{XY}, \widehat{NDE}_{XY} + \widehat{NIE}_{XY})$ is below some threshold.

Definition 4.4 (Completeness)**.** Given threshold δ, reasoning is *complete* w.r.t. $\langle \varphi, \mathcal{M}, \mathcal{Q} \rangle$ when the following holds:

$$\forall\ \varphi_{\mathbf{x}}\ :\ \theta(\varphi^*_{\mathbf{x}}, \widehat{\varphi}_{\mathbf{x}}) \leq \delta \ \text{ and} \tag{10}$$

$$\forall\ \varphi_{\mathbf{x}}, \varphi_{\mathbf{x}'}\ :\ \varphi^*_{\mathbf{x}} = \varphi^*_{\mathbf{x}'} \implies \theta(\widehat{\varphi}_{\mathbf{x}}, \widehat{\varphi}_{\mathbf{x}'}) \leq \delta. \tag{11}$$

Taxonomy of Reasoners Following from Defs. 4.2 and 4.3, we delineate four categories of reasoners:

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

Following from this taxonomy, there are three distinct profiles of incomplete CCR (IC, VI, II) but only one profile that can achieve completeness (VC). To illustrate, consider the logic problem given by Fig. 3:

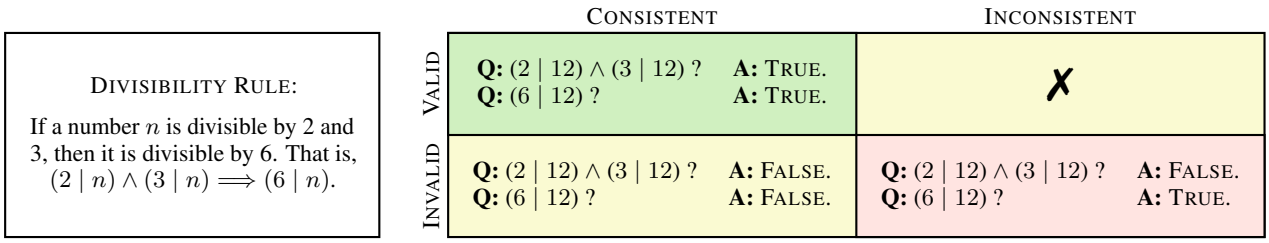


Figure 3. VC, VI, IC, and II reasoning for a logic problem applying divisibility rules. For such a problem, we see that VI reasoning is impossible. However, all four types of reasoners can arise in the probabilistic setting in which we evaluate LMs.

Example 4.5. Consider a logic problem applying the divisibility rule $[(2 \mid n) \wedge (3 \mid n) \implies (6 \mid n)]$ where $n = 12$. A reasoner can respond TRUE or FALSE to each query.

1. **VC**: Each individual query is answered correctly (externally valid) and the logical implication TRUE $\implies$ TRUE is correct (internally consistent).
2. **IC**: Individual queries are answered incorrectly (externally invalid), but the logical implication FALSE $\implies$ FALSE is correct (internally consistent).
3. **II**: One individual response is wrong (externally invalid) and the logical implication FALSE $\implies$ TRUE is incorrect (internally inconsistent).

We see that only VC, IC, and II arise in the logic setting. However, LM evaluation is a probabilistic setting where errors are thresholded. Thus, all four types of reasoners can arise in LM evaluation, as illustrated in Section 6.

Implementation This conceptual introduction has left implementation details intentionally vague. Section 5 suggests one possible procedure for assessing compositional consistency in causal reasoning (Alg. 1), based on compositional properties of the ATE and PNS in graphs with cutpoints. Section 6 provides an empirical illustration of this framework in LMs, demonstrating proof of viability.

5. Inductive CCR Evaluation for the PrC

5.1. PNS Composition Across Graph Components

The PNS boasts several convenient properties for reasoning evaluation: (1) variables of interest are binary and probabilities are bounded by 0 and 1; (2) translating PrC queries to text prompts designed to elicit logical, mathematical, probabilistic, and/or causal reasoning is relatively straightforward (González & Nori, 2024; Hüyük et al., 2025); and (3) the PNS and ATE coincide under certain conditions (Propsition B.4), and thus share convenient compositional properties.

Assumptions: DAGs with Cutpoints In this work, we derive compositional forms for the PrC in graphs with cutpoints. A *cutpoint*, *cut vertex*, or *articulation point* is any node contained in multiple *biconnected components* (BCCs): maximal biconnected subgraphs induced by a partition of edges, such that two edges are in the same partition if and only if they share a common simple cycle (Westbrook & Tarjan, 1992). Thus, removing a cutpoint disconnects the graph. In this evaluation framework, we assume that the causal DAG $\mathcal{G}_{XY}$ representing our SCM contains 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, and (**A4**) no observed nor unobserved confounders for $\{X, Y\}$. Thus, we only consider models with a single "source" node whose causal influence follows multiple indirect pathways to a single "sink" node. See Figs. 4 and D.1 for DAGs that satisfy these assumptions.[4]

PNS Compositionality Though the PN and PS have proven useful for AI reasoning evaluation (González & Nori, 2024; Hüyük et al., 2025), we prove in Appendix B.2 that their composition across BCCs is complex. However, both the PNS and ATE display a simple compositional form under the following conditions. When the DAG $\mathcal{G}_{XY}$ of a linear SCM satisfies A1–A4, the ATE for the root and leaf of $\mathcal{G}_{XY}$ is a product of the ATE values for the root and leaf of each BCC. This follows from the product-of-coefficients heuristic used in classical path-tracing and mediation analysis (Alwin & Hauser 1975; see Appendix C for a worked example). This is not guaranteed for the ATE in nonlinear data generating processes. Conveniently, we prove in Appendix B.2 that this property extends to the PNS under monotonicity even when causal functions are nonlinear.

Theorem 5.1 (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*

$$\mathrm{PNS}_{XY} = \prod_{\{R_i, L_i\} \in \mathbf{C}} \mathrm{PNS}_{R_i L_i} \tag{12}$$

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.

Adjacent BCCs in $\mathcal{G}_{XY}$ can be treated as a single BCC and Theorem 5.1 still holds. In Appendix E, we illustrate

[4] Appendix C explores violations of the cutpoint assumption.

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 $\mathcal{G}_{XY}$.

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 $\widehat{\varphi}^\circ_i$ for path i from knowledge of edges $j \in i$
  $\epsilon_i \leftarrow \theta(\varphi^*_{XY}, \widehat{\varphi}^\circ_i)$ ▷ External validity.
  $\gamma_i \leftarrow \theta(\widehat{\varphi}_{XY}, \widehat{\varphi}^\circ_i)$ ▷ Internal consistency.
return η, ϵ, γ

Theorem 5.1 with simulations of inductive and deductive CCR for the ATE and PNS.

To facilitate CCR evaluation for compositional forms similar to Theorem 5.1, we introduce an intuitive graphical tool for visualizing the flow of causal information through BCCs: the *commutative cut tree*.

Definition 5.2 (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 5.1). 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 $\mathcal{G}_{XY}$ (e.g., $X \to C \to D \to Y$ for Fig. 4A).
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 (e.g., Fig. 4B).

CCTs abstract away complexity in our original causal graph by collapsing BCCs into single edges. This allows for evaluation on arbitrarily complex DAGs with cutpoints as if they were simply directed chains. This abstraction simplifies the problem representation by (1) marginalizing out variables that are unnecessary for valid causal inference in our setting and (2) visualizing pathways of composition. As demonstrated in Section 5.2, CCTs can be leveraged as a design tool for formulating reasoning tasks. As illustrated in Section 6.2, they can also serve as an interpretable, intuitive visualization tool for graphically representing CCR successes and failures.

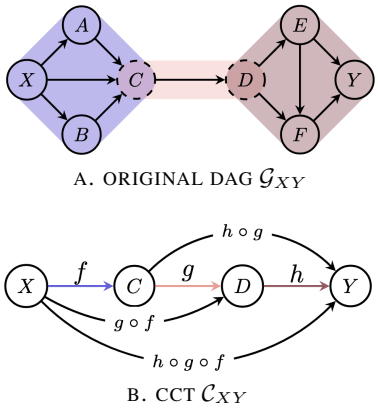


Figure 4. A running example for our framework. (**A**) DAG with BCCs (violet, pink, maroon) and cutpoints C, D. (**B**) CCT $\mathcal{C}_{XY}$ modeling all commutative CCR paths from X to Y. Here, $f := \text{PNS}_{XC}$, $g := \text{PNS}_{CD}$, $h := \text{PNS}_{DY}$, $g \circ f := \text{PNS}_{XD}$, $h \circ g \circ f := \text{PNS}_{XY}$, etc., and composition is multiplicative.

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

Table 1. Quantities of interest for inductive CCR over the PNS for Fig. 4. All compositions are equivalent to the global quantity.

5.2. Inductive CCR as Commutative Reasoning

We now propose a means of systematically evaluating CCR that leverages CCTs and Theorem 5.1 (Alg. 1), applicable to the ATE under linearity and the PNS under monotonicity. We can view CCR as reasoning that $\mathcal{C}_{XY}$ is commutative: every possible composition (corresponding to paths in $\mathcal{C}_{XY}$) should be equivalent to each other and to ground truth, up to some error. When all errors returned by Alg. 1 are less than some threshold(s) δ, the model is complete for task $\langle\varphi, \mathcal{M}, \mathcal{Q}\rangle$. Time complexity is discussed in Appendix D.

Running Example: Intuition for Algorithm 1 For illustration, we walk through Alg. 1 where φ is the PNS and DAG $\mathcal{G}_{XY}$ is structured according to Fig. 4A. We continue to employ the notation defined in Section 4. In Section 6, we implement this same walk-through for LM evaluation.

The assumption stated in Alg. 1 is satisfied, as the PNS composes over BCCs by multiplication (Theorem 5.1). To begin, we must determine the quantities of interest for assessing CCR over $\mathcal{G}_{XY}$ (Table 1). To do this, we first obtain the corresponding CCT $\mathcal{C}_{XY}$. The global quantity of interest will always be the PNS for the root and leaf (PNS_{XY}), corresponding to edge $X \to Y$ in $\mathcal{C}_{XY}$. Local quantities of

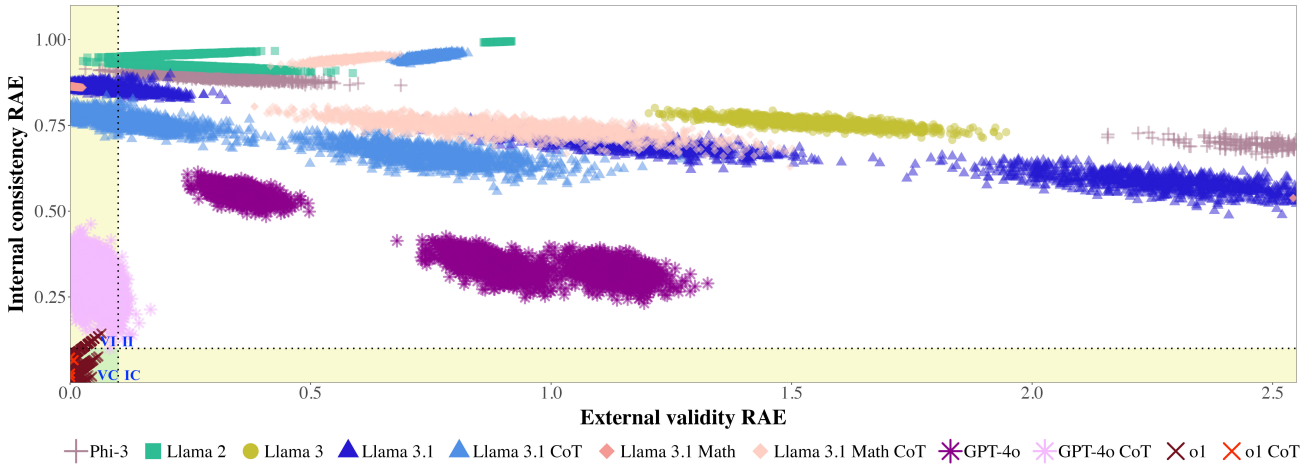


Figure 5. Composition RAE with respect to ground truth (external validity) and $\widehat{\text{PNS}}_{XY}$ (internal consistency). Dotted lines represent the error threshold (RAE = 0.1), with reasoning quadrants VI/IC in yellow, VC in green, and II in white. Models are listed by increasing size (Table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

interest will be the PNS for every remaining pair of nodes in $\mathcal{C}_{XY}$, resulting in $\binom{n}{2} - 1$ quantities (e.g., PNS_{CY}). Finally, we obtain compositions by considering every distinct indirect path from X to Y in $\mathcal{C}_{XY}$. For each such path, the composition of interest is the product of the PNS values for each cause-effect pair on that path. In this case, there are three indirect paths from X to Y: $X \to C \to Y$, $X \to D \to Y$, and $X \to C \to D \to Y$. Thus, our compositions of interest are $\text{PNS}_{XC}\text{PNS}_{CY}$, $\text{PNS}_{XD}\text{PNS}_{DY}$, and $\text{PNS}_{XC}\text{PNS}_{CD}\text{PNS}_{DY}$.

Input to Alg. 1 is the set of estimates $\widehat{\text{PNS}}_{\cdot} = \mathcal{A}(\mathcal{Q}_{\text{PNS}_{\cdot}})$ for each quantity enumerated in Table 1, as well as their ground truth values (if available). Lines 1–2 compute θ for the external validity of the PNS for each cause-effect pair in $\mathcal{C}_{XY}$. Lines 3–6 compute θ for the external validity and internal consistency of each composition. Each $\widehat{\varphi}_i^{\circ}$ corresponds to one distinct path from X to Y in $\mathcal{C}_{XY}$, and is obtained by taking the product of PNS estimates for each cause-effect pair on the path. For example, we estimate $\text{PNS}_{XC}\text{PNS}_{CY}$ as $\widehat{\text{PNS}}_{XC}$ times $\widehat{\text{PNS}}_{CY}$. Finally, Alg. 1 returns all external validity and internal consistency errors.

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.
```

Figure 6. Factual and counterfactual prompt excerpts. For example, we can obtain $\widehat{\text{PNS}}_{XC}$ by simulating potential outcomes $X = \text{TRUE}$, $X = \text{FALSE}$ (Xinyu is or is not happy) and then querying for the value of C (Celine is or is not happy). Analogously, we obtain $\widehat{\text{PNS}}_{DY}$ with interventions on D (Daphne's happiness) and queries on Y (Yasmin's happiness), etc. Responses were converted to booleans using Llama 3. Full prompts in Appendix F.

6. Empirical Demonstration in LMs

We demonstrate Alg. 1 for evaluating inductive CCR for the PNS with an illustrative math problem graphically represented by Fig. 4 (full details in Appendix F). To assess CCR at the counterfactual rung of Pearl's Causal Hierarchy, we treat LMs as *counterfactual data simulators* (González & Nori, 2024) by prompting models with factual and counterfactual queries (e.g., Figs. 6, F.3, F.4).

Note that experiments were designed to demonstrate the implementation of this evaluation framework and to display its usefulness, not to thoroughly assess CCR in current LMs. Thus, we focus deeply on one toy problem to highlight the kinds of insights that our framework can provide. Moreover, success on CCR tasks such as this is *necessary but not sufficient* for demonstrating that LMs can reason.

6.1. Experimental Design

Models Inference was performed using seven architectures: Llama versions 2 (Touvron et al., 2023), 3 (Dubey et al., 2024), 3.1 (Dubey et al., 2024), and 3.1 Math (Tosh-

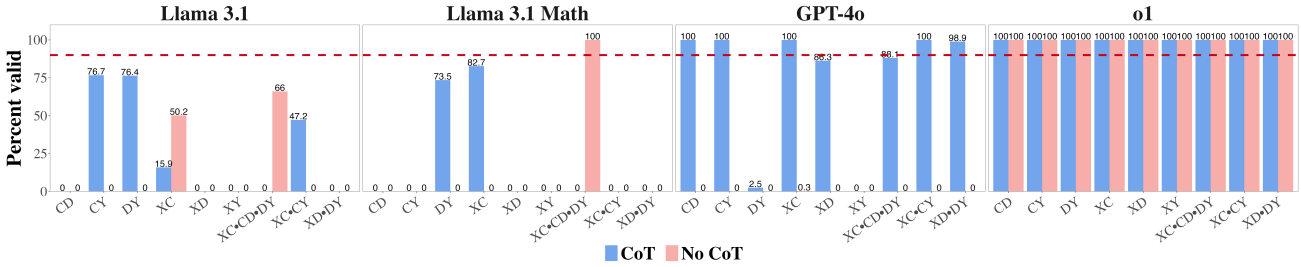


Figure 7. Percent of PNS estimates ($n = 1000$) that are externally valid for CoT vs non-CoT prompting. PNS are denoted by cause-effect pair (e.g., $\text{PNS}_{XC}\text{PNS}_{CY}$ as $XC \cdot CY$). Reasoning was externally valid if $\geq 90\%$ of estimates had RAE ≤ 0.1 (red dashed line).

niwal et al., 2024); Phi-3-mini (Abdin et al., 2024); GPT-4o (Achiam et al., 2023); and o1 (Jaech et al., 2024) (Table F.1). All LMs were fine-tuned for dialogue. Llama 3.1 Math was also fine-tuned for math reasoning, and o1 is a designated reasoning model.

CCR Task Let $\mathcal{M}$ be an SCM represented by the DAG in Fig. 4. Variables $\mathbf{V} \in \mathcal{M}$ are binary and causal functions are logical *or* (Eq. F.1). Quantities of interest are in Table 1. We defined CCR task $\mathcal{T} := \langle \text{PNS}, \mathcal{M}, \mathcal{Q} \rangle$ where prompts $\mathcal{Q}$ were based on the `CandyParty` math word problem (Figs. 6, F.1; González & Nori 2024). Approximation errors were obtained by computing relative absolute errors (RAE; Eq. F.3). Thus, we define a task-metric-model triple $\langle \mathcal{T}, \text{RAE}, \mathcal{A} \rangle$ for each model $\mathcal{A}$. For Llama 3.1, Llama 3.1 Math, and GPT-4o, we also presented a CoT formulation of $\mathcal{Q}$ to assess impacts of CoT on CCR elicitation. A wrapper for our original prompt template used two examples to demonstrate CoT for the model: one factual and one counterfactual (Fig. F.2). All other details of the CoT experiment were consistent with the non-CoT design.

Extracting and Evaluating PNS Values For each quantity of interest, we sampled 1000 sets of exogenous variable values, providing PNS distributions rather than point estimates. For each set, we generated one factual and one counterfactual problem using the text prompt template (Fig. F.1). Five answers were sampled for each problem. The corresponding boolean value was extracted from text responses using Llama 3. See Appendix F for full details on boolean extraction. Reasoning was considered externally valid for a quantity if $\geq 90\%$ of estimates had RAE ≤ 0.1 (threshold chosen prior to analysis). Reasoning was deemed near-valid if $\geq 75\%$ of estimates met the threshold.

6.2. Experimental Results

General Results Fig. F.7 plots all PNS distributions. Figs. 5 and F.6 jointly represent internal consistency and external validity RAE for our three compositions of interest, with quadrants representing reasoning profiles: VC, VI, IC, II. Figs. 7, F.8, and F.9 show the percent of PNS estimates that were externally valid for each model. Errors did not decrease monotonically with increasing model size nor Llama version (Figs. F.11, F.12). Preliminary analyses suggest that poor numeracy might explain some errors (Figs. F.3, F.4).

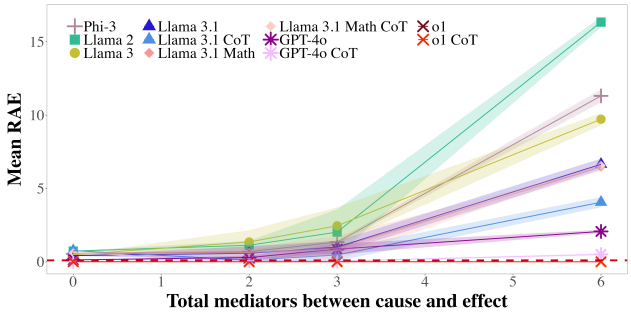


Figure 8. RAE increases with total mediating variables for most models. Red dashed line denotes external validity cutoff (RAE = 0.1), with standard deviations in shaded regions.

Taxonomy of Reasoners Results for task $\mathcal{T}$ revealed three taxonomically distinct error patterns: VC (o1), near-VI (GPT-4o with CoT), and II (all remaining models). Only o1 – the lone "reasoning" model evaluated – placed mass in the VC reasoning quadrant (Fig. 5). All models besides o1 were 0% valid for global quantity PNS_{XY}, with and without CoT (Figs. 7, F.8, F.9). Without CoT, Phi-3, Llama 2, Llama 3, and GPT-4o failed to exceed 3% validity for any quantity. II reasoners often displayed high variance (Fig. 5). GPT-4o benefited the most from CoT with factual and counterfactual examples, achieving external validity for five of nine quantities and near-validity for two additional. However, internal inconsistency made GPT-4o with CoT a near-VI reasoner: 75.1% of all estimates were near-valid, mean external validity on local quantities and compositions exceeded 77% and 95%, respectively, and yet failure to correctly infer PNS_{XY} revealed internally inconsistent reasoning.

Visualizing Reasoning Pathways with CCTs As demonstrated in Fig. 9, CCTs allow us to see valid and invalid compositional reasoning pathways in graphical representation. Paths from X to Y highlighted in red represent

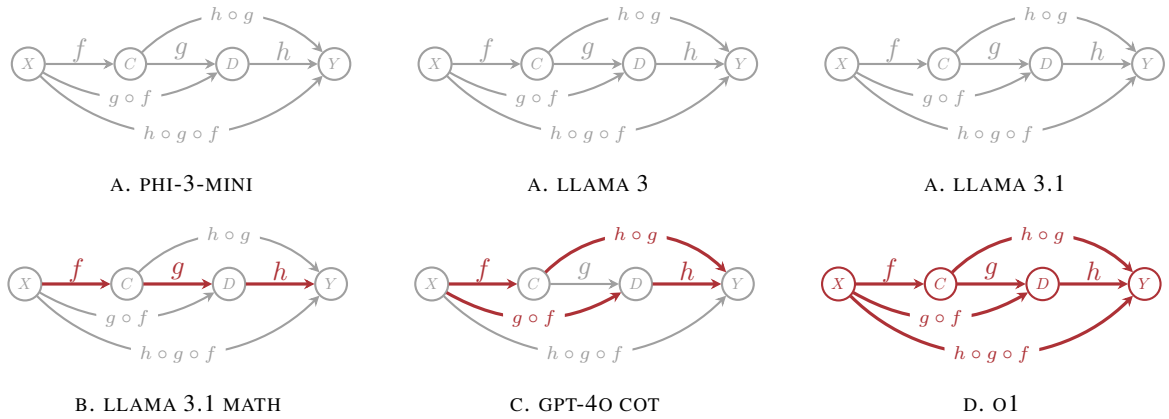


Figure 9. Visualizing valid compositional reasoning pathways with CCTs. Each red path from X to Y represents externally valid compositions, while gray represents invalid reasoning. Externally valid local estimates are not shown. Nodes are red when all paths passing through them are valid. Best results between CoT and non-CoT are pictured for each model (defaulting to non-CoT when equivalent). Results for Llama 2 are equivalent to Llamas 3 and 3.1.

externally valid compositions. Only o1 was a VC (complete) reasoner for task $\mathcal{T}$, as illustrated by a fully red CCT. Moving from Llama 3.1 Math to GPT-4o (CoT) to o1, we see progressively more reasoning pathways are highlighted (1, 2, and all paths from X to Y, respectively).

Together with quantitative results (e.g., Figs. 5 and 7), these qualitative visualizations highlight the nuanced view provided by our framework: considering compositional consistency in the form of commutative reasoning adds extra dimensions of informativeness relative to evaluating on quantity-wise accuracy alone. With our reasoning taxonomy, differing patterns of incorrectness emerge.

Errors Increase With Mediation Errors were assessed w.r.t. the complexity of paths between cause and effect. For all models except GPT-4o with CoT and o1, mean RAE increased monotonically as shortest path distance and total mediating variables increased (Figs. 8, F.13). Nevertheless, mean RAE for GPT-4o with CoT was over $10\times$ higher for six mediators than for three mediators. Increasing RAE may be explained by error propagation. Additionally, preliminary analyses suggests that models can "lose the train of thought" along long causal paths (Fig. F.5). These results are in line with prior findings that reasoning performance can diminish with task complexity (Agrawal et al., 2017; Ma et al., 2023; Ray et al., 2023; Press et al., 2023; Jin et al., 2023).

7. Limitations & Future Directions

This work presents a conceptual foundation for CCR evaluation in LMs. We instantiate this framework for the ATE and PNS in causal graphs with cutpoints (Alg. 1). Preliminary empirical results are limited to one illustrative task as proof of viability. Results demonstrate how this framework can provide a richer picture of causal reasoning performance than measuring quantity-wise external validity alone, as is conventionally done (González & Nori, 2024). For large-scale benchmarking, future work could explore the automated design of inductive and deductive CCR tasks with varying graphical complexity. As this work only considers the ATE and PNS under Theorem 5.1, extensions could consider other estimands and compositional forms.

Impact Statement

The potential for reasoning emergence in AI has broad scientific, economic, and social implications for global society. These implications include matters of safety and fairness. This work aims to promote the rigorous measurement of reasoning behavior in LMs. Success on CCR tasks such as those described in this work is *necessary but not sufficient* for demonstrating that LMs can reason, and we encourage cautious interpretation of results.

References

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agrawal, A., Kembhavi, A., Batra, D., and Parikh, D. C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. *arXiv preprint arXiv:1704.08243*, 2017.

Alwin, D. F. and Hauser, R. M. The decomposition of effects in path analysis. *American sociological review*, pp. 37–47, 1975.

Avin, C., Shpitser, I., and Pearl, J. Identifiability of path-specific effects. In *IJCAI International Joint Conference on Artificial Intelligence*, pp. 357–363, 2005.

Cai, H., Wang, Y., Jordan, M., and Song, R. On Learning Necessary and Sufficient Causal Graphs, November 2023. URL http://arxiv.org/abs/2301.12389. arXiv:2301.12389 [cs, stat].

Coecke, B. Compositionality as we see it, everywhere around us. In *The Quantum-Like Revolution: A Festschrift for Andrei Khrennikov*, pp. 247–267. Springer, 2023.

De Toffoli, S. 'chasing'the diagram—the use of visualizations in algebraic reasoning. *The Review of Symbolic Logic*, 10(1):158–186, 2017.

Du, L., Ding, X., Xiong, K., Liu, T., and Qin, B. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 432–446, 2022.

Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., and Grathwohl, W. S. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International Conference on Machine Learning*, pp. 8489–8510. PMLR, 2023.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., Welleck, S., West, P., Bhagavatula, C., Le Bras, R., et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36, 2023.

Fong, B. and Spivak, D. I. *An invitation to applied category theory: seven sketches in compositionality*. Cambridge University Press, 2019.

Frankland, S. M. and Greene, J. D. Concepts and compositionality: in search of the brain's language of thought. *Annual review of psychology*, 71(1):273–303, 2020.

Goddu, M. K. and Gopnik, A. The development of human causal learning and reasoning. *Nature Reviews Psychology*, pp. 1–21, 2024.

González, J. and Nori, A. V. Does reasoning emerge? examining the probabilities of causation in large language models. In *Advances in Neural Information Processing Systems*, 2024.

Haugeland, J. Understanding natural language. *The Journal of Philosophy*, 76(11):619–632, 1979.

He, X., Tian, Y., Sun, Y., Chawla, N. V., Laurent, T., LeCun, Y., Bresson, X., and Hooi, B. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *arXiv preprint arXiv:2402.07630*, 2024.

Hsieh, C.-Y., Zhang, J., Ma, Z., Kembhavi, A., and Krishna, R. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 37, 2023.

Huang, J. and Chang, K. C.-C. Towards reasoning in large language models: A survey. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67: 757–795, 2020.

Hüyük, A., Xu, X., Maasch, J., Nori, A. V., and González, J. Reasoning elicitation in language models via counterfactual feedback. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.03767.

Imai, K., Keele, L., and Tingley, D. A general approach to causal mediation analysis. *Psychological methods*, 15(4): 309, 2010.

Imbens, G. W. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Ito, T., Klinger, T., Schultz, D., Murray, J., Cole, M., and Rigotti, M. Compositional generalization through abstract representations in human and artificial neural networks. *Advances in Neural Information Processing Systems*, 35:32225–32239, 2022.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Jeong, H., Tian, J., and Bareinboim, E. Finding and listing front-door adjustment sets. *Advances in Neural Information Processing Systems*, 35:33173–33185, 2022.

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Zhiheng, L., Blin, K., Adauto, F. G., Kleiman-Weiner, M., Sachan, M., et al. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh conference on neural information processing systems*, 2023.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Kschischang, F. R., Frey, B. J., and Loeliger, H.-A. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory*, 47(2):498–519, 2001.

Lake, B. M. and Baroni, M. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Li, A. and Pearl, J. Probabilities of Causation with Nonbinary Treatment and Effect. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):20465–20472, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i18.30030. URL https://ojs.aaai.org/index.php/AAAI/article/view/30030.

Li, X. L., Shrivastava, V., Li, S., Hashimoto, T., and Liang, P. Benchmarking and improving generator-validator consistency of language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Li, Y., Sreenivasan, K., Giannou, A., Papailiopoulos, D., and Oymak, S. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.-W., Wu, Y. N., Zhu, S.-C., and Gao, J. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36, 2023a.

Lu, P., Qiu, L., Yu, W., Welleck, S., and Chang, K.-W. A survey of deep learning for mathematical reasoning. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14605–14631, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.817. URL https://aclanthology.org/2023.acl-long.817.

Ma, Z., Hong, J., Gul, M. O., Gandhi, M., Gao, I., and Krishna, R. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

MacKinnon, D. *Introduction to statistical mediation analysis*. Routledge, 2012.

Maiti, A., Plecko, D., and Bareinboim, E. Counterfactual identification under monotonicity constraints. In *The 39th Annual AAAI Conference on Artificial Intelligence*, 2025.

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

Mueller, S., Li, A., and Pearl, J. Causes of effects: Learning individual responses from population data. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, 2022.

Okawa, M., Lubana, E. S., Dick, R., and Tanaka, H. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Advances in Neural Information Processing Systems*, 36, 2023.

Pearl, J. Reverend bayes on inference engines: A distributed hierarchical approach. *Proceedings, AAAI-82*, pp. 133–136, 1982.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Pearl, J. Probabilities of Causation: Three Counterfactual Interpretations and Their Identification. *Synthese*, 121: 93–149, 1999.

Pearl, J. Causality: Models, reasoning, and inference. *Cambridge, UK: Cambridge University Press*, 19(2):3, 2000.

Pearl, J. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, 2001.

Pearl, J. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96, 2009.

Pearl, J. The mediation formula: A guide to the assessment of causal pathways in nonlinear models. *Causality: Statistical perspectives and applications*, pp. 151–179, 2012.

Pearl, J. Linear Models: A Useful "Microscope" for Causal Analysis. *Journal of Causal Inference*, 1(1):155–170, 2013.

Pearl, J. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014.

Plečko, D. and Bareinboim, E. Causal fairness analysis: A causal toolkit for fair machine learning. 17(3):304–589, 2024. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000106. URL http://www.nowpublishers.com/article/Details/MAL-106.

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., and Lewis, M. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., and Chen, H. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5368–5393, 2023.

Qiu, L., Jiang, L., Lu, X., Sclar, M., Pyatkin, V., Bhagavatula, C., Wang, B., Kim, Y., Choi, Y., Dziri, N., et al. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. In *The Twelfth International Conference on Learning Representations*, 2024.

Ray, A., Radenovic, F., Dubey, A., Plummer, B. A., Krishna, R., and Saenko, K. Cola: A benchmark for compositional text-to-image retrieval. *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.03689.

Sani, N. and Mastakouri, A. A. Tightening Bounds on Probabilities of Causation By Merging Datasets, October 2023. URL http://arxiv.org/abs/2310.08406. arXiv:2310.08406 [cs, math, stat].

Sani, N., Mastakouri, A. A., and Janzing, D. Bounding probabilities of causation through the causal marginal problem, April 2023. URL http://arxiv.org/abs/2304.02023. arXiv:2304.02023 [cs, math, stat].

Saxton, D., Grefenstette, E., Hill, F., and Kohli, P. Analysing mathematical reasoning abilities of neural models. In *International Conference on Learning Representations*, 2019.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage? *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., and Behrens, T. Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell*, 186(22):4885–4897, 2023.

Shafer, G. R. and Shenoy, P. P. Probability propagation. *Annals of Mathematics and Artificial Intelligence*, 2:327–351, 1990.

Shi, Z., Liu, H., Min, M. R., Malon, C., Li, L. E., and Zhu, X. Retrieval, analogy, and composition: A framework for compositional generalization in image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1990–2000, 2021.

Singal, R. and Michailidis, G. Axiomatic effect propagation in structural causal models. *Journal of Machine Learning Research*, 25(52):1–71, 2024.

Stolfo, A., Jin, Z., Shridhar, K., Schoelkopf, B., and Sachan, M. 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)*, pp. 545–561, 2023.

Stone, A., Wang, H., Stark, M., Liu, Y., Scott Phoenix, D., and George, D. Teaching compositionality to cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5058–5067, 2017.

Su, J., Liu, N., Wang, Y., Tenenbaum, J. B., and Du, Y. Compositional image decomposition with diffusion models. *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Suhr, A., Lewis, M., Yeh, J., and Artzi, Y. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Tian, J. and Pearl, J. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

Toshniwal, S., Du, W., Moshkov, I., Kisacanin, B., Ayrapetyan, A., and Gitman, I. OpenMathInstruct-2: Accelerating AI for math with massive open-source instruction data, 2024. URL https://arxiv.org/abs/2410.01560.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

VanderWeele, T. J. A unification of mediation and interaction: a 4-way decomposition. *Epidemiology*, 25(5): 749–761, 2014.

VanderWeele, T. J. Mediation analysis: a practitioner's guide. *Annual review of public health*, 37:17–32, 2016.

Wang, H., Feng, S., He, T., Tan, Z., Han, X., and Tsvetkov, Y. Can language models solve graph problems in natural language? *Advances in Neural Information Processing Systems*, 36, 2023a.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023b.

Wang, Y. and Jordan, M. I. Desiderata for Representation Learning: A Causal Perspective, February 2022. URL http://arxiv.org/abs/2109.03795. arXiv:2109.03795 [cs, stat].

Wang, Z., Do, Q. V., Zhang, H., Zhang, J., Wang, W., Fang, T., Song, Y., Wong, G., and See, S. Cola: Contextualized commonsense causal reasoning from the causal inference perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5253–5271, 2023c.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Westbrook, J. and Tarjan, R. E. Maintaining bridge-connected and biconnected components on-line. *Algorithmica*, 7(1):433–464, 1992.

Wittgenstein, L. *Philosophical investigations*. John Wiley & Sons, 2009.

Wright, S. Correlation and causation. *Journal of agricultural research*, 20(7):557, 1921.

Wu, A., Brantley, K., and Artzi, Y. A surprising failure? multimodal llms and the nlvr challenge. *arXiv preprint arXiv:2402.17793*, 2024.

Wüst, A., Stammer, W., Delfosse, Q., Dhami, D. S., and Kersting, K. Pix2code: Learning to compose neural visual concepts as programs. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Xu, Z., Niethammer, M., and Raffel, C. A. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. *Advances in Neural Information Processing Systems*, 35:25074–25087, 2022.

Yang, L., Shirvaikar, V., Clivio, O., and Falck, F. A critical review of causal reasoning benchmarks for large language models. In *AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?"*, 2024.

Zečević, M., Willig, M., Dhami, D. S., and Kersting, K. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023.

Zhang, J. and Bareinboim, E. Non-parametric path analysis in structural causal models. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

Zhang, W., Wu, T., Wang, Y., Cai, Y., and Cai, H. Towards trustworthy explanation: On causal rationalization. In *International Conference on Machine Learning*, pp. 41715–41736. PMLR, 2023.

APPENDIX

A. EXTENDED BACKGROUND

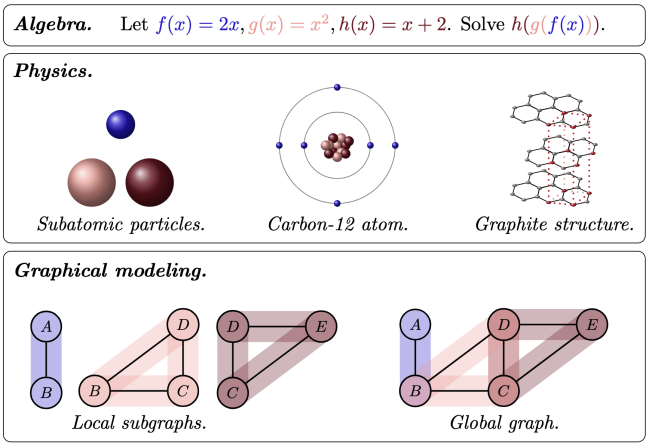


Figure A.1. Composition is ubiquitous in physical and symbolic systems, operating at multiple scales of granularity: atoms → molecules → materials; cells → tissues → organs; words → sentences; subroutines → programs; substructures → global graphs; etc.

Compositional Reasoning in Generative AI Compositional reasoning has been explored for diverse tasks in generative AI. These include problem-solving in math, logic, and programming (Saxton et al., 2019; Dziri et al., 2023; Lu et al., 2023a); image generation (Du et al., 2023; Okawa et al., 2023), decomposition (Su et al., 2024), and captioning (Shi et al., 2021); visual concept learning (Wüst et al., 2024); object recognition (Stone et al., 2017); representation learning (Xu et al., 2022); chain-of-thought (Li et al., 2023); visual question answering (Agrawal et al., 2017); and text-to-image retrieval (Ray et al., 2023). Models have been shown to struggle with compositional reasoning in vision and language, especially for complex compositions (Agrawal et al., 2017; Ma et al., 2023; Ray et al., 2023). Saxton et al. (2019) address compositional math reasoning with an emphasis on compositions of functions and manipulation of intermediate results, though causal measures are not explored. Many works center on the development of benchmark datasets for vision and/or language models (e.g., Johnson et al. 2017; Agrawal et al. 2017; Suhr et al. 2017; 2018; Thrush et al. 2022; Ma et al. 2023; Hsieh et al. 2023; Ray et al. 2023). This work introduces new core concepts in CCR evaluation for LMs, on which future benchmarks can be based.

Applications of the PrC in AI The majority of works on the PrC have centered on identifiability and the derivation of bounds under various conditions (Tian & Pearl, 2000; Mueller et al., 2022; Sani et al., 2023; Sani & Mastakouri, 2023; Li & Pearl, 2024; Maiti et al., 2025). The PrC have found applications in representation learning (Wang & Jordan, 2022), causal discovery (Cai et al., 2023), and explainable AI and rationalization (Zhang et al., 2023). In LMs, the PrC have been used to evaluate causal reasoning (González & Nori, 2024) and to facilitate fine-tuning for counterfactual reasoning (Hüyük et al., 2025). To our knowledge, ours is the first work to explore the utility of PrC compositionality for measuring compositional reasoning in AI.

Commutative Diagrams in Reasoning Evaluation Classically, commutative diagrams have been used as tools for theorem proving (see *diagram chasing*, De Toffoli 2017). In this work, we introduce a special form of commutative diagram for modeling reasoning. González & Nori (2024) use commutative diagrams for causal reasoning evaluation in LMs, where one pathway corresponds to the true problem solution for a math problem and the other corresponds to the reasoning pathway of the LM. The present work shows how commutative diagrams in the form of CCTs can provide compact and exhaustive representations for (1) visualizing CCR pathways and (2) systematizing CCR evaluation.

B. PROBABILITIES OF CAUSATION: COMPOSITIONALITY

B.1. PN & PS

The PS considers the *presence* of a causal process that can produce a given effect, while the PN considers the *absence* of alternative explanatory processes (Tian & Pearl, 2000). Like the PNS, the PN and PS are point identifiable when Y is monotonic in X.

Definition B.1 (Probability of Necessity (PN), Pearl 1999)**.** The probability that event y would *not* have occurred in the absence of event x, given that x and y did jointly occur, is given as

$$\begin{aligned} PN &\coloneqq \mathbb{P}(y'_{x'}|x,y) \\ &= \mathbb{P}(Y_{x'} = \text{FALSE}|X = \text{TRUE}, Y = \text{TRUE}). \end{aligned} \tag{B.1}$$

Definition B.2 (Probability of Sufficiency (PS), Pearl 1999)**.** The probability that event x would produce event y given that x and y did *not* in fact occur is given as

$$PS \coloneqq \mathbb{P}(y_x|x',y'). \tag{B.2}$$

Definition B.3 (Point identification of the PN and PS under monotonicity, Tian & Pearl 2000)**.** Given a positive-Markovian SCM for which Y is monotonic in X, the causal effects $\mathbb{P}(y_x)$ and $\mathbb{P}(y_{x'})$ are identifiable by Equation 1 and the probabilities of causation are point identifiable by the following expressions.

$$PN = \frac{\mathbb{P}(y) - \mathbb{P}(y_{x'})}{\mathbb{P}(x,y)} = \frac{\mathbb{P}(y) - \mathbb{P}(y|do(x'))}{\mathbb{P}(x,y)} \tag{B.3}$$

$$PS = \frac{\mathbb{P}(y_x) - \mathbb{P}(y)}{\mathbb{P}(x',y')} = \frac{\mathbb{P}(y|do(x)) - \mathbb{P}(y)}{\mathbb{P}(x',y')}. \tag{B.4}$$

B.2. PrC Compositionality: Proofs

Proof. Here we show that the PN and PS do not compose according to Theorem 5.1, while the PNS does. Consider a causal model $X \to Y \to Z$ with binary variables X, Y, Z. Assume that all relations in the causal graph are monotonic such that

$$\mathbb{P}(y'_x, y_{x'}) = 0 \tag{B.5}$$

$$\mathbb{P}(z'_y, z_{y'}) = 0. \tag{B.6}$$

We start by expressing PN_{XY}, PS_{XY} in terms of probabilities over potential outcomes $Y_x, Y_{x'}$:

$$PN_{XY} = \mathbb{P}(y'_{x'}|x,y) = \frac{\mathbb{P}(y, y'_{x'}|x)}{\mathbb{P}(y, y_{x'}|x) + \mathbb{P}(y, y'_{x'}|x)} = \frac{\mathbb{P}(y_x, y'_{x'})}{\mathbb{P}(y_x, y_{x'}) + \mathbb{P}(y_x, y'_{x'})} \tag{B.7}$$

$$PS_{XY} = \mathbb{P}(y_x|x',y') = \frac{\mathbb{P}(y_x, y'|x')}{\mathbb{P}(y_x, y'|x') + \mathbb{P}(y'_x, y'|x')} = \frac{\mathbb{P}(y_x, y'_{x'})}{\mathbb{P}(y_x, y'_{x'}) + \mathbb{P}(y'_x, y'_{x'})}. \tag{B.8}$$

Using these expressions, together with the monotonicity assumption $\mathbb{P}(y'_x, y_{x'}) = 0$ and the simple fact that

$$\mathbb{P}(y_x, y_{x'}) + \mathbb{P}(y_x, y'_{x'}) + \mathbb{P}(y'_x, y'_{x'}) + \mathbb{P}(y'_x, y_{x'}) = 1, \tag{B.9}$$

we can fully describe the distribution of potential outcomes $Y_x, Y_{x'}$ (by solving the resulting system of equations):

$$\mathbb{P}(y_x, y_{x'}) = \frac{1/PN_{XY} - 1}{1/PN_{XY} + 1/PS_{XY} - 1} \tag{B.10}$$

$$\mathbb{P}(y_x, y'_{x'}) = \frac{1}{1/PN_{XY} + 1/PS_{XY} - 1} \tag{B.11}$$

$$\mathbb{P}(y'_x, y'_{x'}) = \frac{1/PS_{XY} - 1}{1/PN_{XY} + 1/PS_{XY} - 1} \tag{B.12}$$

$$\mathbb{P}(y'_x, y_{x'}) = 0. \tag{B.13}$$

Similarly, given PN_{YZ}, PS_{YZ}, we describe the distribution of potential outcomes $Z_y, Z_{y'}$:

$$\mathbb{P}(z_y, z_{y'}) = \frac{1/PN_{YZ} - 1}{1/PN_{YZ} + 1/PS_{YZ} - 1} \tag{B.14}$$

$$\mathbb{P}(z_y, z'_{y'}) = \frac{1}{1/PN_{YZ} + 1/PS_{YZ} - 1} \tag{B.15}$$

$$\mathbb{P}(z'_y, z'_{y'}) = \frac{1/PS_{YZ} - 1}{1/PN_{YZ} + 1/PS_{YZ} - 1} \tag{B.16}$$

$$\mathbb{P}(z'_y, z_{y'}) = 0. \tag{B.17}$$

Since we have now fully characterized our causal model, we can easily derive formulas for PN_{XZ} or PS_{XZ}. Starting with PN_{XZ}, we first observe that

$$PN_{XZ} = \frac{\mathbb{P}(z_x, z'_{x'})}{\mathbb{P}(z_x, z_{x'}) + \mathbb{P}(z_x, z'_{x'})}, \tag{B.18}$$

as we have done with PN_{XY} earlier. Then, we note that

$$\begin{aligned}
\mathbb{P}(z_x, z'_{x'}) &= \mathbb{P}(y_x, y_{x'})\mathbb{P}(z_x, z'_{x'}|y_x, y_{x'}) + \mathbb{P}(y_x, y'_{x'})\mathbb{P}(z_x, z'_{x'}|y_x, y'_{x'}) + \mathbb{P}(y'_x, y'_{x'})\mathbb{P}(z_x, z'_{x'}|y'_x, y'_{x'}) && \text{(B.19)}\\
&= \mathbb{P}(y_x, y_{x'})\mathbb{P}(z_y, z'_y) + \mathbb{P}(y_x, y'_{x'})\mathbb{P}(z_y, z'_{y'}) + \mathbb{P}(y'_x, y'_{x'})\mathbb{P}(z_{y'}, z'_{y'}) && \text{(B.20)}\\
&= \mathbb{P}(y_x, y'_{x'})\mathbb{P}(z_y, z'_{y'}) && \text{(B.21)}\\
&= \frac{1}{1/PN_{XY} + 1/PS_{XY} - 1} \times \frac{1}{1/PN_{YZ} + 1/PS_{YZ} - 1} && \text{(B.22)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(z_x, z_{x'}) &= \mathbb{P}(y_x, y_{x'})\mathbb{P}(z_x, z_{x'}|y_x, y_{x'}) + \mathbb{P}(y_x, y'_{x'})\mathbb{P}(z_x, z_{x'}|y_x, y'_{x'}) + \mathbb{P}(y'_x, y'_{x'})\mathbb{P}(z_x, z_{x'}|y'_x, y'_{x'}) && \text{(B.23)}\\
&= \mathbb{P}(y_x, y_{x'})\mathbb{P}(z_y, z_y) + \mathbb{P}(y_x, y'_{x'})\mathbb{P}(z_y, z_{y'}) + \mathbb{P}(y'_x, y'_{x'})\mathbb{P}(z_{y'}, z_{y'}) && \text{(B.24)}\\
&= \mathbb{P}(y_x, y_{x'})\mathbb{P}(z_y) + \mathbb{P}(y_x, y'_{x'})\mathbb{P}(z_y, z_{y'}) + \mathbb{P}(y'_x, y'_{x'})\mathbb{P}(z_{y'}) && \text{(B.25)}\\
&= \mathbb{P}(y_x, y_{x'})(\mathbb{P}(z_y, z_{y'}) + \mathbb{P}(z_y, z'_{y'})) + \mathbb{P}(y_x, y'_{x'})\mathbb{P}(z_y, z_{y'}) + \mathbb{P}(y'_x, y'_{x'})(\mathbb{P}(z_y, z_{y'}) + \mathbb{P}(z'_y, z_{y'})) && \text{(B.26)}\\
&= \frac{1}{1/PN_{XY} + 1/PS_{XY} - 1} \times \frac{1}{1/PN_{YZ} + 1/PS_{YZ} - 1} \times \\
&\qquad \left[\left(\frac{1}{PN_{XY}} - 1\right)\left(\frac{1}{PN_{YZ}}\right) + \left(\frac{1}{PN_{YZ}} - 1\right) + \left(\frac{1}{PS_{XY}} - 1\right)\left(\frac{1}{PN_{YZ}} - 1\right)\right]. && \text{(B.27)}
\end{aligned}$$

Therefore,

$$\frac{1}{PN_{XZ}} = \frac{1}{PN_{XY}}\frac{1}{PN_{YZ}} + \left(\frac{1}{PS_{XY}} - 1\right)\left(\frac{1}{PN_{YZ}} - 1\right), \tag{B.28}$$

which cannot be reduced further to a function of PN_{XY} and PN_{YZ} only. This is because our causal model has enough degrees of freedom to fix PN_{XY} and PN_{YZ} but still vary PS_{XY} resulting in different values of PN_{XZ}.

Following a similar derivation, we can also write

$$\frac{1}{PS_{XZ}} = \frac{1}{PS_{XY}}\frac{1}{PS_{YZ}} + \left(\frac{1}{PN_{XY}} - 1\right)\left(\frac{1}{PS_{YZ}} - 1\right). \tag{B.29}$$

Thus, while individual PN and PS values compose in a rather complicated way, PNS values happen to compose multiplicatively. We have already proven this statement along the way in (B.21):

$$PNS_{XZ} = \mathbb{P}(z_x, z'_{x'}) = \mathbb{P}(y_x, y'_{x'})\mathbb{P}(z_y, z'_{y'}) = PNS_{XY}PNS_{YZ} \tag{B.30}$$

Thus, the PNS composes according to Theorem 5.1 under monotonicity. To obtain (B.21), we used monotonicity in (B.19). Otherwise, there would have been a fourth term in (B.19) related to $\mathbb{P}(y'_x, y_{x'})$ and we would have been left with

$$\mathbb{P}(z_x, z'_{x'}) = \mathbb{P}(y_x, y'_{x'})\mathbb{P}(z_y, z'_{y'}) + \mathbb{P}(y'_x, y_{x'})\mathbb{P}(z'_y, z_{y'}). \tag{B.31}$$

This is intuitive: if $X \iff Z$, then either $X \iff Y \wedge Y \iff Z$ or $X \iff Y' \wedge Y' \iff Z$. Then, you can argue that PNS_{XZ} cannot be a pure function of PNS_{XY} and PNS_{YZ}: we can fix $PNS_{XY} = \mathbb{P}(y_x, y'_{x'})$ and $PNS_{YZ} = \mathbb{P}(z_y, z'_{y'})$ and still easily vary $\mathbb{P}(y'_x, y_{x'})$ or $\mathbb{P}(z'_y, z_{y'})$ resulting in different values of PNS_{XZ}.

□

B.3. Extended Discussion of the PNS

B.3.1. The PNS Coincides with the ATE Under Monotonicity

We highlight the relationship of the PNS to another causal quantity: the *average treatment effect* (ATE). Under monotonicity, the PNS is identifiable from causal effects. In the binary setting, the causal effect $\mathbb{E}[Y \mid do(X = x)]$ can be expressed as

$$\mathbb{E}[Y \mid do(X = x)] = 1 \cdot \mathbb{P}(y = 1 \mid do(X = x)) + 0 \cdot \mathbb{P}(y = 0 \mid do(X = x)) \tag{B.32}$$
$$= \mathbb{P}(y = 1 \mid do(X = x)). \tag{B.33}$$

Proposition B.4. *Following from Eq. B.33, we can express the ATE as a difference of probabilities that is equal to the PNS.*

$$\text{ATE} := \mathbb{E}[Y \mid do(X = x)] - \mathbb{E}[Y \mid do(X = x')] \tag{B.34}$$
$$= \mathbb{P}(y \mid do(X = x)) - \mathbb{P}(y \mid do(X = x')) \tag{B.35}$$
$$= PNS. \tag{B.36}$$

C. ATE Composition in Linear SCMs

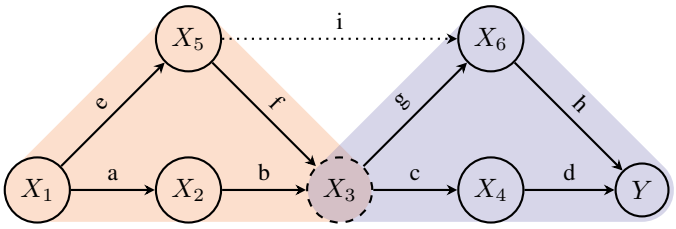


Figure C.1. DAG $\mathcal{G}_{X_1Y}$ contains a subgraph that can be decomposed into two BCCs sharing cutpoint X_3. $\mathcal{G}_{X_1X_3}$ is the BCC induced by edges in orange, $\mathcal{G}_{X_3Y}$ by edges in periwinkle. Assume a linear SCM. If the dotted edge from X_5 to X_6 does not exist, $\text{ATE}_{X_1Y} = \text{ATE}_{X_1X_3} \cdot \text{ATE}_{X_3Y}$. If the dotted edge does exist, then this product is summed with an additional term corresponding to the path-specific effect for path $X_1 \to X_5 \to X_6 \to Y$, which does not pass through X_3.

C.1. Worked Example

In linear SCMs, composition of the ATE over the BCCs of a DAG shares the same form as Theorem 5.1. This follows from the product-of-coefficients heuristic used in classical path-tracing and linear mediation analysis (Wright, 1921; Alwin & Hauser, 1975; MacKinnon, 2012; Imai et al., 2010; Pearl, 2012; 2013; Singal & Michailidis, 2024). Linear path-tracing rules are not guaranteed to apply in nonlinear data generating processes, as addressed in nonparametric causal mediation analysis (Pearl, 2001; 2012) and nonparametric path-tracing (Zhang & Bareinboim, 2018).

We provide an illustrative example of composition for the ATE in linear SCMs for two settings: (1) when the DAG contains at least one cutpoint and (2) when the DAG does not contain a cutpoint, but does contain a subgraph with at least one cutpoint. Finite sample results are in Figure C.2.

DAGs With Cutpoints Take Figure C.1 as an example. First, we assume that the dotted edge from X_5 to X_6 does not exist. In this case, the DAG contains cutpoint X_3 and two BCCs. The ATE for $\{X_1, Y\}$ can then be expressed as a product

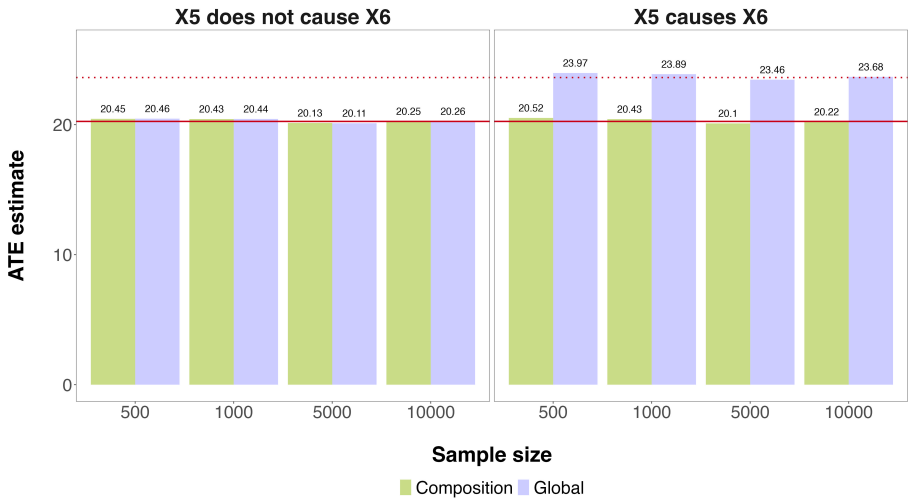


Figure C.2. ATE composition for a linear SCM whose graphical representation is given by Figure C.1. Composition = $\text{ATE}_{X_1X_3} \cdot \text{ATE}_{X_3Y}$ and global = ATE_{X_1Y}. The solid red line represents the true ATE_{X_1Y} when $X_5 \not\to X_6$, and the dotted red line represents ATE_{X_1Y} when $X_5 \to X_6$. Estimates were obtained by linear regression (https://scikit-learn.org/).

of the ATE values corresponding to the root and leaf of each BCC.

$$\text{ATE}_{X_1X_3} = ab + ef \tag{C.1}$$
$$\text{ATE}_{X_3Y} = cd + gh \tag{C.2}$$
$$\text{ATE}_{X_1Y} = abcd + abgh + efcd + efgh \tag{C.3}$$
$$\text{ATE}_{X_1X_3} \cdot \text{ATE}_{X_3,Y} = (ab + ef)(cd + gh) = \text{ATE}_{X_1Y} \tag{C.4}$$

We can see that the ground truth $\text{ATE}_{X_1,Y}$ expressed in Equation C.3 is equivalent to the product expressed in Equation C.4.

DAGs Without Cutpoints Next, assume that the edge from X_5 to X_6 does exist.

$$\text{ATE}_{X_1Y} = abcd + abgh + efcd + efgh + eih \tag{C.5}$$
$$= (ab + ef)(cd + gh) + eih \tag{C.6}$$
$$= \text{ATE}_{X_1X_3} \cdot \text{ATE}_{X_3Y} + \text{PSE} \tag{C.7}$$

where PSE is the path-specific effect for $X_1 \to X_5 \to X_6 \to Y$, the only causal path for $\{X_1, Y\}$ not passing through cut vertex X_3.

Finite Sample Simulation In Figure C.2, we demonstrate finite sample results for the case where $X_5 \to X_6$ and $X_5 \not\to X_6$. Exogenous variables are drawn from the standard normal distribution. All edge coefficients in Figure C.1 are set to 1.5. Thus, we have the following true ATE values when $X_5 \not\to X_6$:

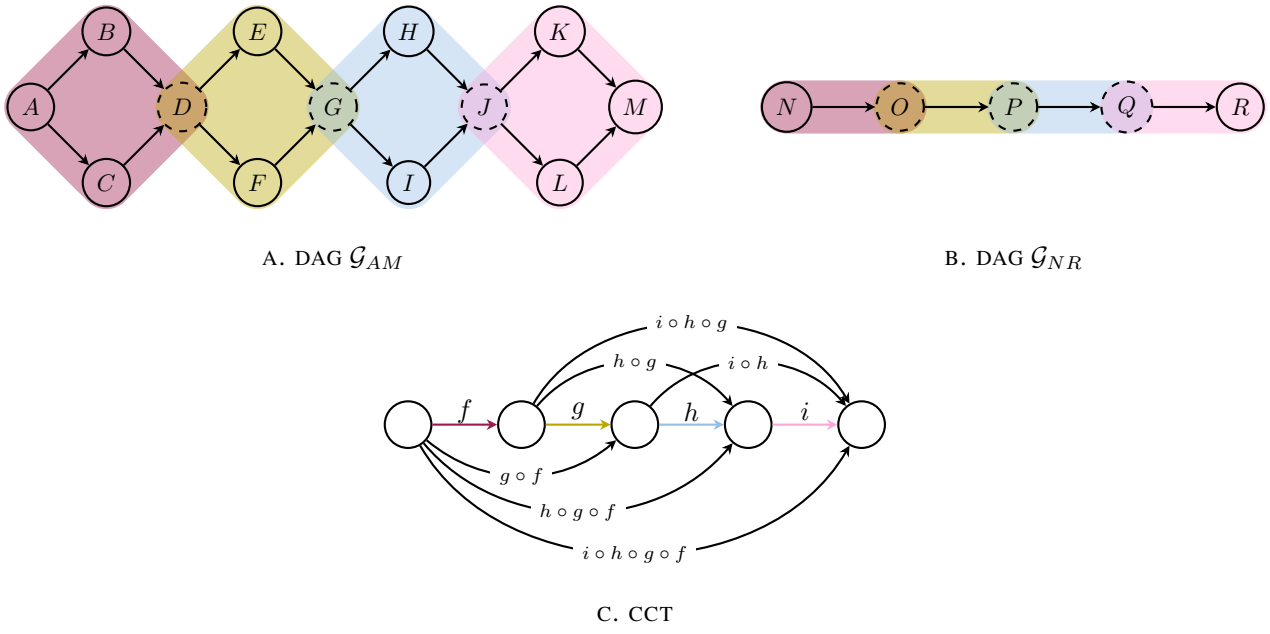
$$\text{ATE}_{X_13} = 1.5^2 + 1.5^2 = 4.5 \tag{C.8}$$
$$\text{ATE}_{X_3Y} = 1.5^2 + 1.5^2 = 4.5 \tag{C.9}$$
$$\text{ATE}_{X_1Y} = 4(1.5^4) = 20.25. \tag{C.10}$$

When $X_5 \to X_6$, we have the additional path-specific effect for $X_1 \to X_5 \to X_6 \to Y$, which is equal to $1.5^3 = 3.375$. Thus,

$$\text{ATE}_{X_1Y} = 20.25 + 3.375 = 23.625. \tag{C.11}$$

As shown in Figure C.2, finite sample results approach the true values with small errors. Note that estimation of ATE_{X_3Y} when $X_5 \to X_6$ requires covariate adjustment for X_5, which acts as a confounder for X_3, Y in this case.

D. ALGORITHM 1: INDUCTIVE CCR EVALUATION



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

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

C. CCT

Figure D.1. (**A**) DAG $\mathcal{G}_{AM}$, whose undirected skeleton is a cactus graph with three cutpoints (dashed nodes). (**B**) DAG $\mathcal{G}_{NR}$, a directed chain. (**C**) The CCT shared by both $\mathcal{G}_{AM}$ and $\mathcal{G}_{NR}$, which models all CCR pathways from root to leaf.

D.1. Time Complexity

Let n be the number of nodes in the original causal DAG. In the worst case, the number of nodes in the corresponding CCT will be n (e.g., Figure D.1B). The first for-loop in Algorithm 1 (Lines 2-3) requires errors to be computed for every pair of nodes in the CCT, resulting in $\binom{n}{2}$ errors. Thus, the first for-loop requires the calculation of $O(n^2)$ errors. Let k be the number of unique paths from root to leaf in the CCT. The second for-loop in Algorithm 1 (Lines 3-6) requires two errors to be computed for all $(k-1)$ compositions: one for internal consistency and one for external validation. This requires $O(k)$ errors to be calculated. Thus, the total number of errors to be calculated will be $O(n^2 + k)$. We defer finer-grained analyses to future work.

E. SIMULATIONS

E.1. Numerical Behavior in Finite Samples

The following experiments demonstrate the numerical behavior of inductive and deductive CCR for (1) the PNS when all cause-effect pairs satisfy the monotonicty assumption and (2) the ATE in linear SCMs. All simulations were performed on a MacBook Pro (Apple M2 Pro).

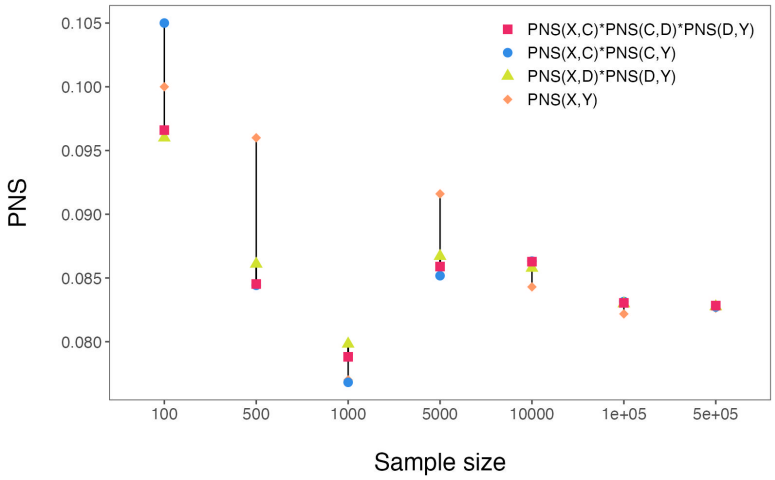


Figure E.1. Finite sample simulations of internally consistent inductive CCR for the PNS. Causal functions were logical *or*. For the CCT in Figure 4, all compositions converged to the same value as sample size increased. See Figure E.3 for analogous results for the ATE in linear-Gaussian SCMs.

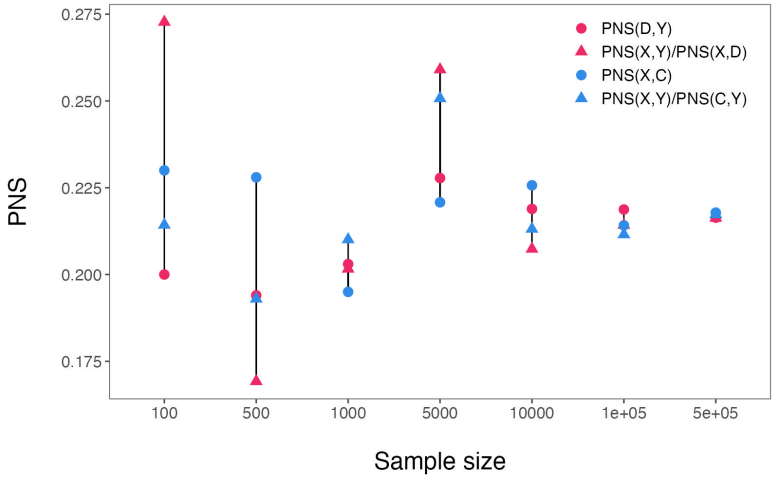


Figure E.2. Finite sample simulations of internally consistent deductive CCR for the PNS. Causal functions were logical *and*. For the CCT in Figure 4, points of the same color were expected to converge. See Figure E.4 for analogous results for the ATE in linear-Gaussian SCMs.

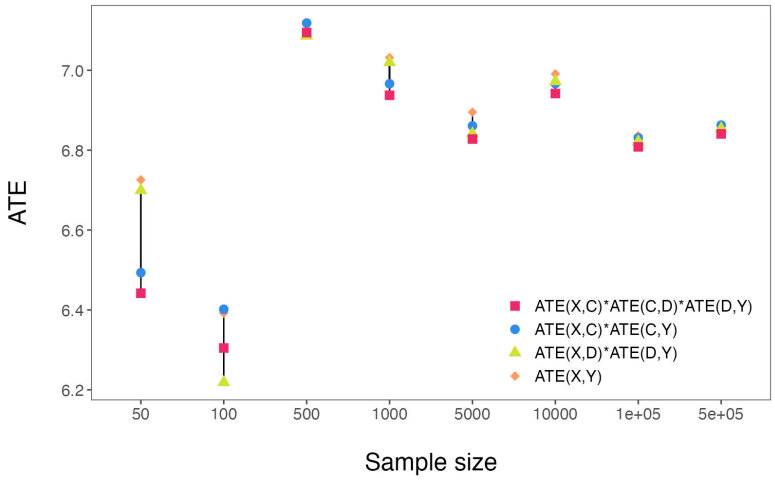


Figure E.3. Finite sample simulations of internally consistent inductive CCR for the ATE in linear-Gaussian SCMs. For the CCT in Figure 4, compositions converged as sample size increased. Estimates were obtained by linear regression (https://scikit-learn.org/).

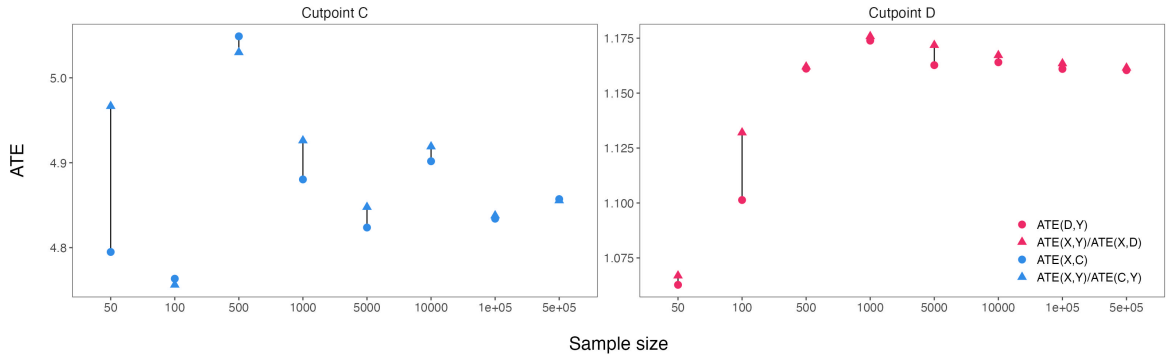


Figure E.4. Finite sample simulations of internally consistent deductive CCR for the ATE in linear-Gaussian SCMs. For the CCT in Figure 4, compositions converged as sample size increased. Estimates were obtained by linear regression (https://scikit-learn.org/).

F. LM EXPERIMENTS

F.1. Experimental Design

Translating Causal Queries to Text We translated the data generating process represented by the DAG in Figure 4 to a mathematical word problem, as expressed in the prompt in Figure F.1. This prompt is based on the `CandyParty` problem described in González & Nori (2024) and Hüyük et al. (2025). Additionally, we implement a CoT wrapper for our original prompt template that uses two examples that demonstrate CoT for the model: one factual and one counterfactual (Figure F.2). Questions and answers used in the CoT scenario were sampled and calculated identically to the non-CoT experiments.

We define SCM $\mathcal{M} \coloneqq \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, p(\mathbf{u}) \rangle$, where $\mathbf{V}$ are binary variables representing the nodes in Figure 4, causal functions $f \in \mathcal{F}$ are logical *or* ($\vee$), and distribution $p(\mathbf{u})$ is Bernoulli. Logical *or* is a monotone boolean function, satisfying the monotonicity condition for point identifiability of the PNS from causal effects (Definition 2.5).

Each node of the ground truth graph is treated as a person in our word problem: X = Xinyu, A = Ara, B = Becca, C = Celine, D = Daphne, E = Emma, F = Fox, Y = Yasmin. All $V_i \in \mathbf{V}$ are given by

$$V_i = pa_1, \vee ... \vee pa_k \vee \text{Ber}(0.1\, T_{V_i}) \tag{F.1}$$

where $\{pa_j\}_{j=1}^{k}$ are the k parents of V_i. All $\{T \cdot\}$ in the context prompt take a value of 7, such that all exogenous variables are drawn from Bernoulli distributions parameterized by $p = 0.7$.

Examples of factual and counterfactual questions and responses are given in Figures F.3 and F.4. For counterfactual questions, a new assumption was introduced about the variable acting as the cause. The model was asked about the variable acting as the effect.

Extracting PNS Values from Text To compute the PNS, model text responses were translated to the corresponding boolean answer. The corresponding boolean was extracted using Llama 3, given the following prompt: "I will give you a question and its answer. Determine whether the meaning of the answer is 'TRUE' or 'FALSE'. An answer is 'TRUE' if it contains phrases like 'yes', 'it holds', 'correct', 'true', or similar affirmations. An answer is 'FALSE' if it contains phrases like 'no', 'it does not hold', 'incorrect', 'false', or similar negations. Respond only with one word: 'TRUE' or 'FALSE'. Question: 'q' Answer: 'a' Is the meaning 'TRUE' or 'FALSE'?"

Model Inference We used a single A100 GPU for all experiments. Models used for inference were LMs fine-tuned for dialogue, as reported in Table F.1. LLama 3.1 Math was also fine-tuned for math reasoning. OpenAI's o1 model is marketed as a "high intelligence reasoning model."[5] We did not perform any additional fine-tuning. We will release the inference code in the final version of this paper.

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).

Approximation Errors Metrics for computing approximation errors were relative absolute error (RAE), which were computed for each estimate ($n = 1000$ PNS estimates per cause-effect pair). The RAE is the absolute error (AE) normalized by the true PNS. When comparing errors across different quantities (e.g., PNS_{XY} versus PNS_{XC}), normalization is needed. Thus, we generally only report the RAE. For external validity, these metrics are computed with respect to ground

[5] https://platform.openai.com/docs/guides/reasoning

truth ($PNS^{*}\cdot$).

$$AE_{\text{external}} := \mid PNS^{*}\cdot - \widehat{PNS}\cdot \mid \tag{F.2}$$

$$RAE_{\text{external}} := \frac{\mid PNS^{*}\cdot - \widehat{PNS}\cdot \mid}{PNS^{*}\cdot} \tag{F.3}$$

For internal consistency, these metrics are computed using estimates for two equivalent quantities ($\widehat{PNS}\cdot$ and $\widehat{PNS}\cdot'$).

$$AE_{\text{internal}} := \mid \widehat{PNS}\cdot - \widehat{PNS}\cdot' \mid \tag{F.4}$$

$$RAE_{\text{internal}} := \frac{\mid \widehat{PNS}\cdot - \widehat{PNS}\cdot' \mid}{\widehat{PNS}\cdot} \tag{F.5}$$

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 {TX} candies.  Ara will be happy if Xinyu is happy or if he gets at least {TA}
candies.  Becca will be happy if Xinyu is happy or if she gets at least {TB} 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 {TC} candies.  Daphne will be happy if Celine is happy or if she
gets at least {TD} candies.  Emma will be happy if Daphne is happy or if she gets at
least {TE} candies.  Fox will be happy if Daphne is happy or if Emma is happy or if
he gets at least {TF} candies.  Yasmin will be happy if Emma is happy or if Fox is
happy or if she gets at least {TY} candies.  After distributing the candies, Xinyu
gets {NX}, Ara gets {NA}, Becca gets {NB}, Celine gets {NC}, Daphne gets {ND}, Emma
gets {NE}, Fox gets {NF}, and Yasmin gets {NY}."
```

Figure F.1. Context prompt template for inductive CCR evaluation. For all experiments reported in this work, {T·} = 7.

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 3 candies." "Ara will be happy if Xinyu is happy or if he gets at least
4 candies." "Becca will be happy if Xinyu is happy or if she gets at least 10
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." "Daphne will be happy if Celine is happy
or if she gets at least 5 candies." "Emma will be happy if Daphne is happy or if she
gets at least 6 candies." "Fox will be happy if Daphne is happy or if Emma is happy
or if he gets at least 10 candies." "Yasmin will be happy if Emma is happy or if Fox
is happy or if she gets at least 1 candies." "After distributing the candies, Xinyu
gets 6, Ara gets 7, Becca gets 5, Celine gets 1, Daphne gets 3, Emma gets 4, Fox
gets 9, and Yasmin gets 8." "Now, suppose that Xinyu is not happy regardless of the
candy distribution." "Is Becca happy?  Be as concise as possible." "ANSWER:" "Since
Xinyu gets 6 candies, which is more than 3, she is happy." "However, we are asked
to assume Xinyu is not happy regardless." "In that case, Becca is happy only if she
gets at least 10 candies." "Becca gets 5 candies, which is less than 10." "Therefore,
no, Becca is not happy!"
```

Figure F.2. CoT formulation of the original `CandyParty` prompt, where two examples that demonstrate CoT are provided to the model: one factual and one counterfactual.

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.

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.

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.*

F.2. Results

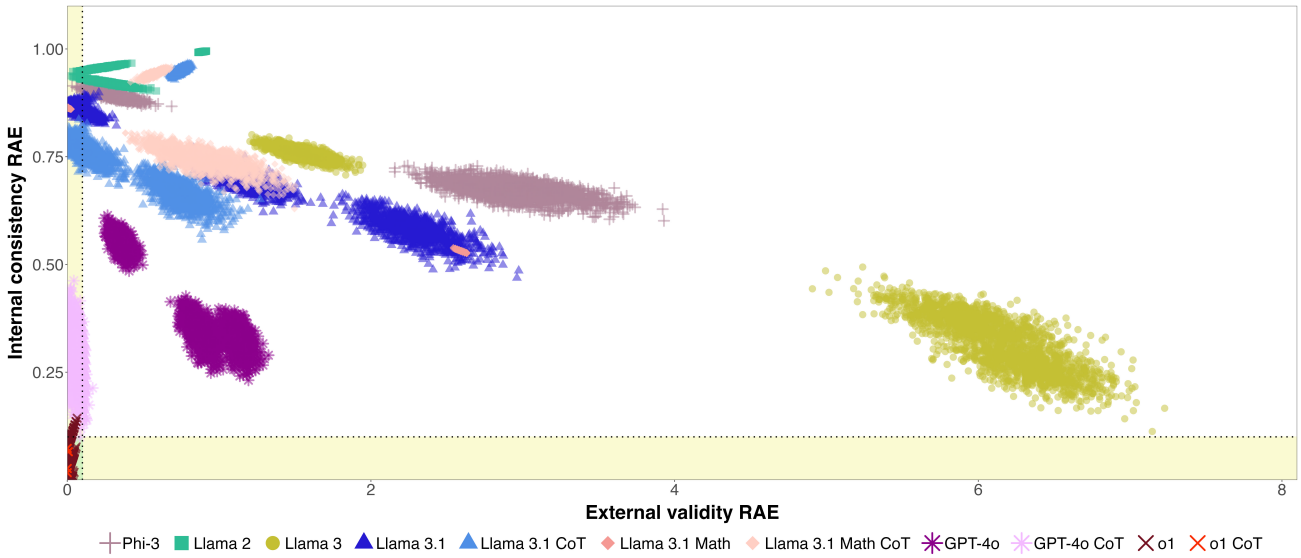


Figure F.6. Composition RAE with respect to ground truth (external validity) and $\widehat{PNS}_{XY}$ (internal consistency) with full x-axis shown. Dotted lines represent the error threshold (RAE = 0.1), with reasoning quadrants VI/IC in yellow, VC in green, and II in white. Models are listed by increasing size.

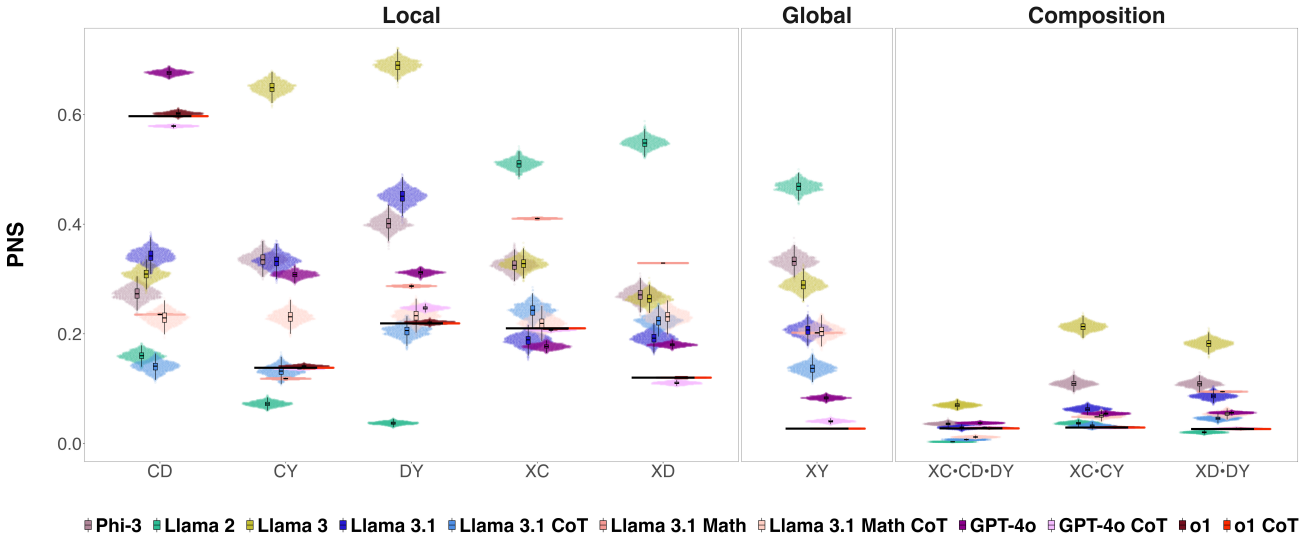


Figure F.7. Estimated PNS distributions ($n = 1000$) for the quantities given in Table 1, denoted here by cause-effect pair (e.g., $XC \cdot CD \cdot DY$ denotes $PNS_{XC}PNS_{CD}PNS_{DY}$). Bold black line segments represent ground truth values.

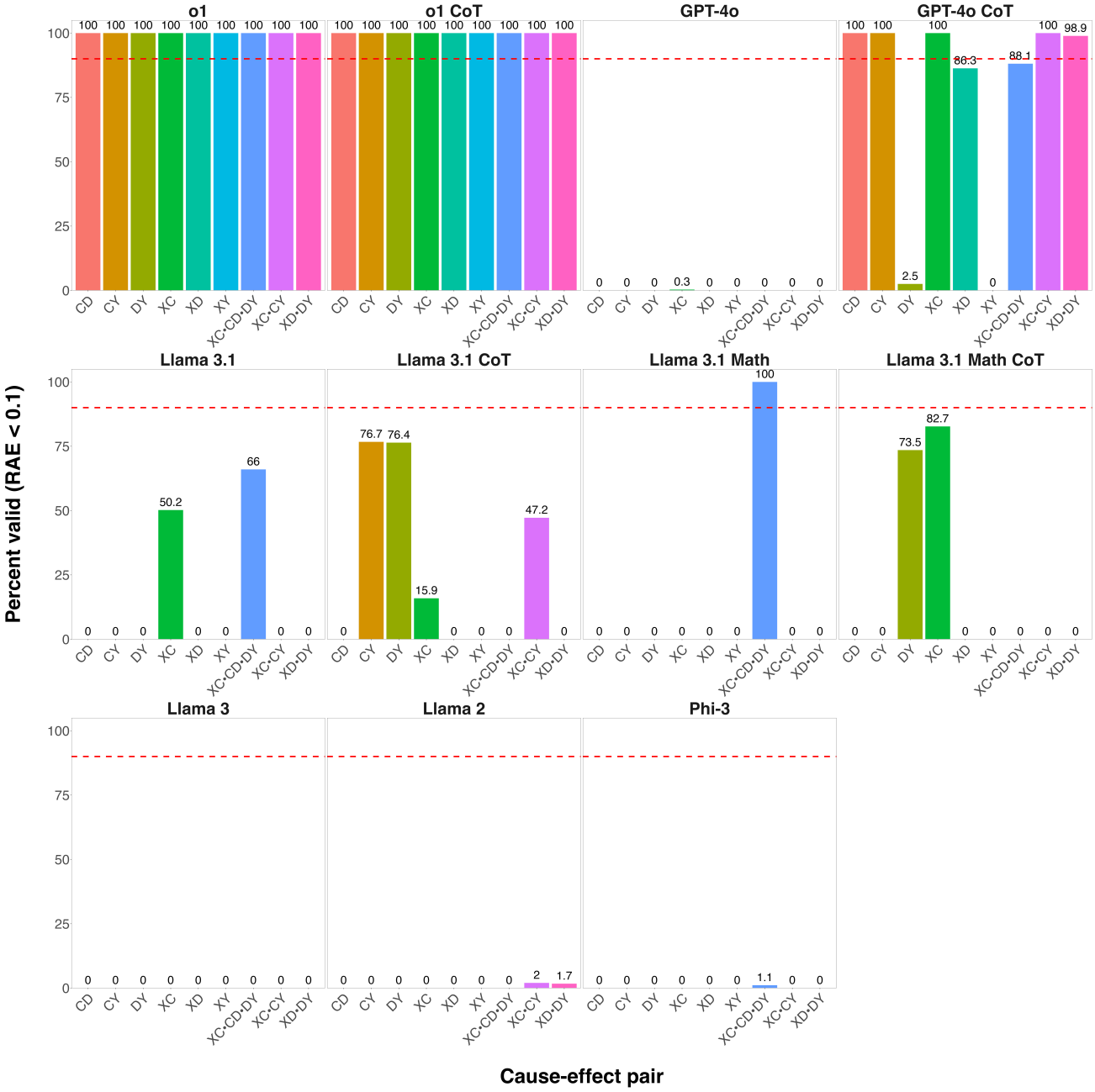


Figure F.8. Percent of PNS estimates ($n = 1000$) that are externally valid. Reasoning was considered externally valid for a quantity if $\geq 90\%$ of estimates had RAE ≤ 0.1 (represented by the red dashed line). PNS are denoted by cause-effect pair (e.g., $XC \cdot CD \cdot DY$ denotes $PNS_{XC}PNS_{CD}PNS_{DY}$).

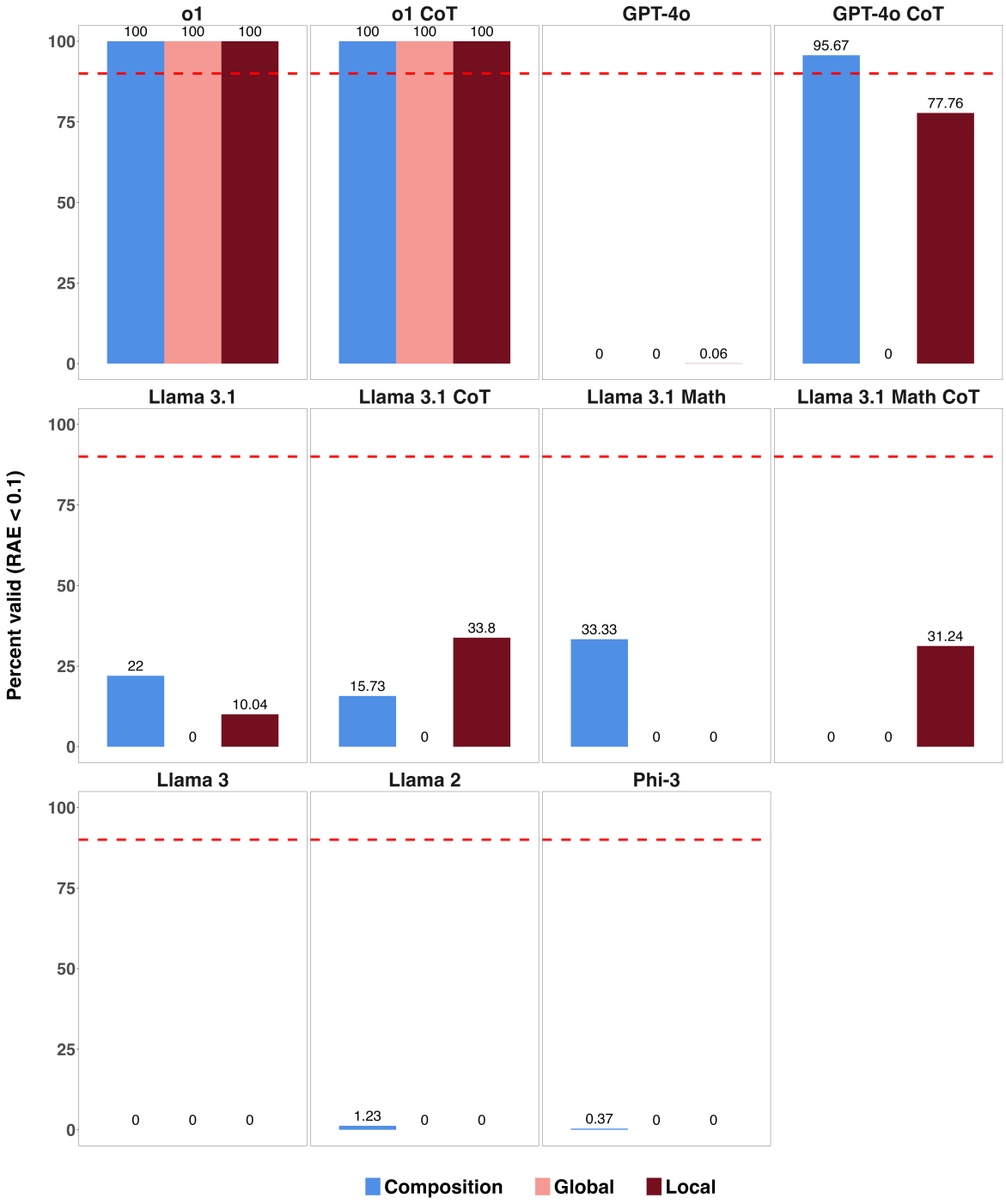


Figure F.9. Percent of PNS estimates ($n = 1000$) that are externally valid. Reasoning was considered externally valid for a quantity if $\geq 90\%$ of estimates had RAE ≤ 0.1 (represented by the red dashed line).

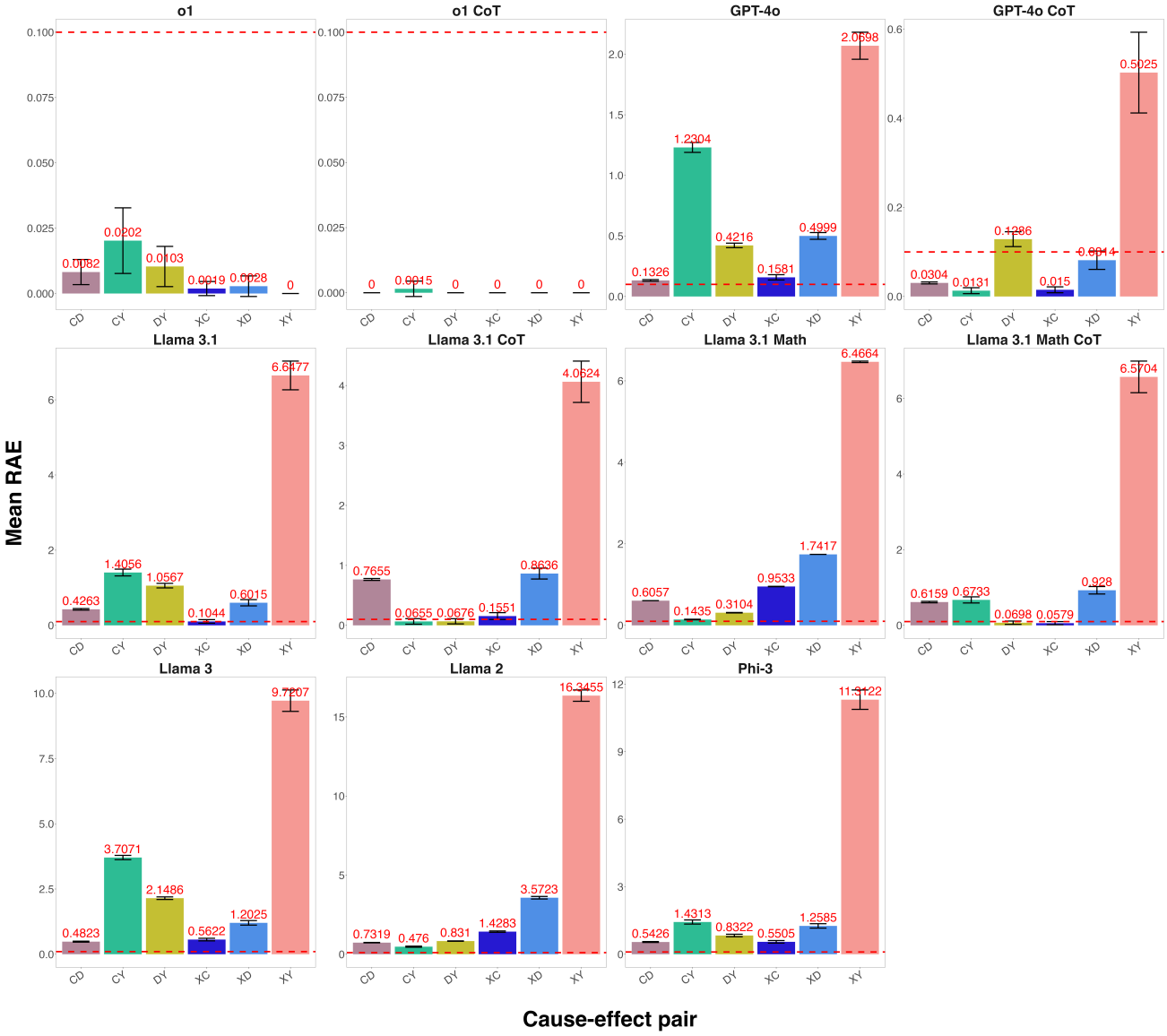


Figure F.10. Mean external validity RAE ($n = 1000$). Error bars represent standard deviations. An estimate was considered externally valid for a quantity if RAE ≤ 0.1 (represented by the red dashed line).

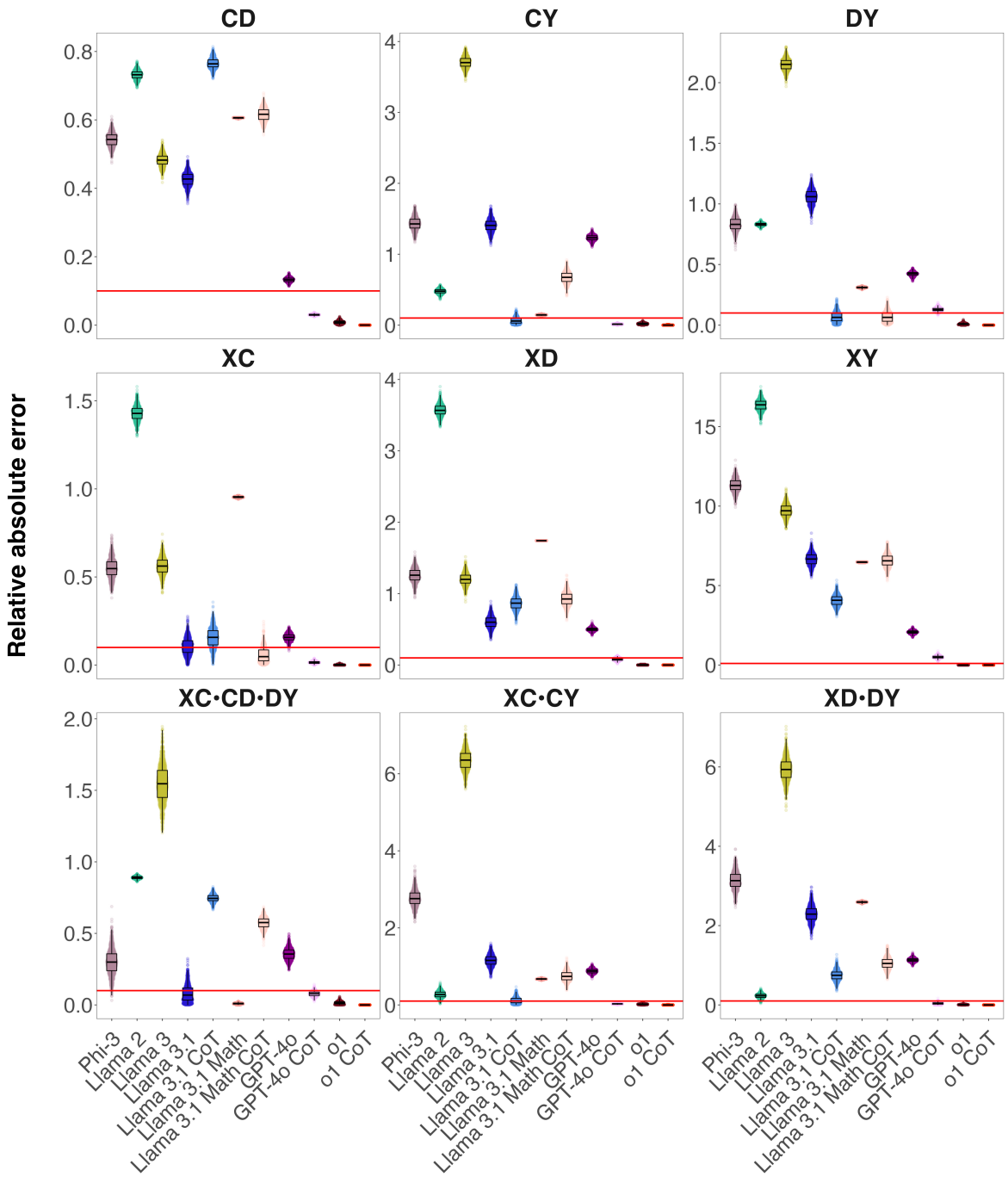


Figure F.11. RAE distributions for all quantities. Red lines represent the external validity cutoff (RAE = 0.1). PNS are denoted by cause-effect pair (e.g., $XC \cdot CD \cdot DY$ denotes $PNS_{XC}PNS_{CD}PNS_{DY}$).

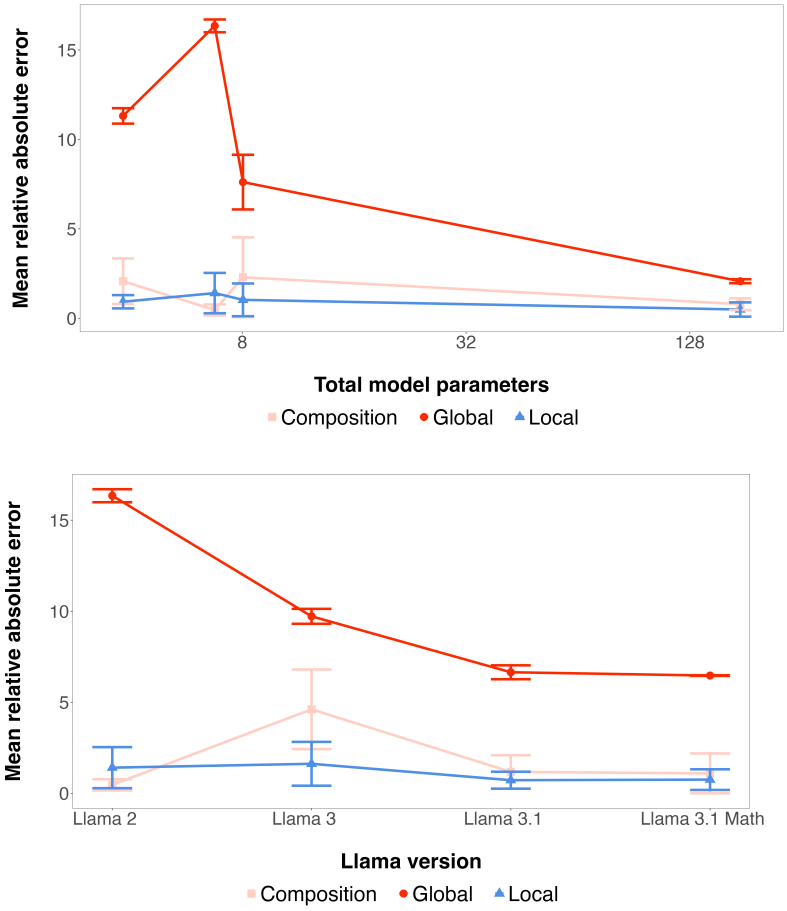


Figure F.12. Mean RAE (without CoT) does not consistently decrease with increasing model size ($\log_2$ scale; **left**) nor Llama version (**right**). However, errors for global quantities (PNS_{XY}) do monotonically decrease with increasing Llama version. Values were averaged separately for global, local, and composed quantities (Table 1). Error bars represent standard deviations. Parameter count for GPT-4o is set to the GPT-3 count (175B) due to information availability.

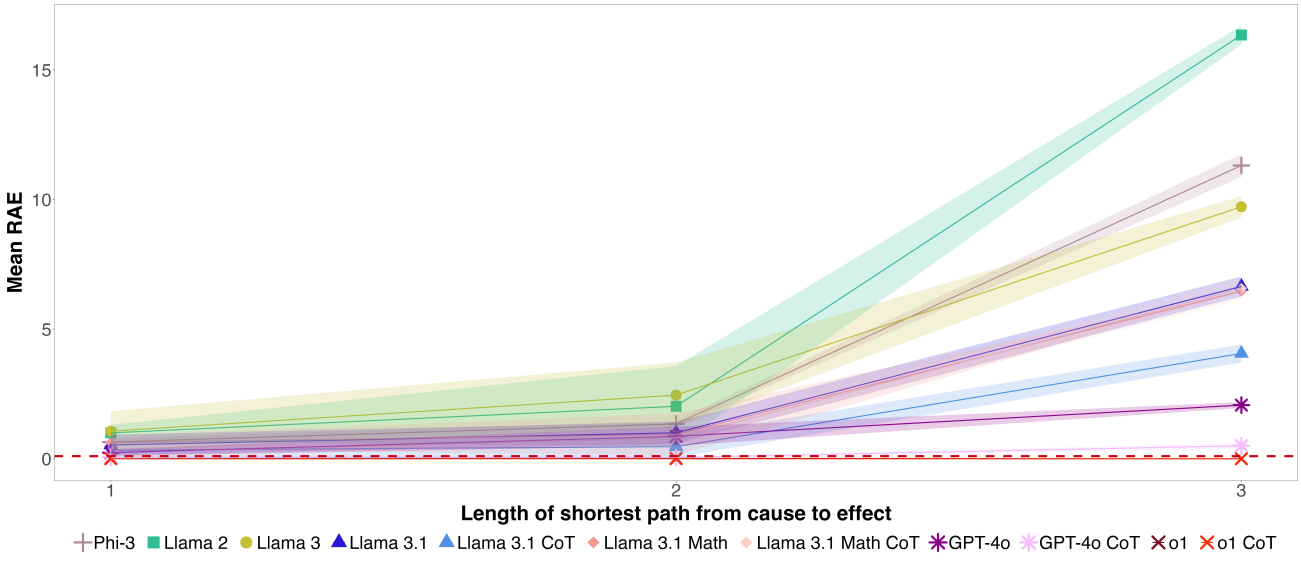


Figure F.13. RAE generally increases with length of shortest path from cause to effect. Red dashed line denotes external validity cutoff (RAE = 0.1), with standard deviations in shaded regions.