Compositional Causal Reasoning Evaluation in Language Models

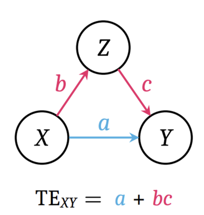
Jacqueline Maasch[‡ ✉], Alihan Hüyük[†], Xinnuo Xu[††], Aditya V. Nori[††], Javier Gonzalez[††] | [‡]Cornell Tech, [†]Harvard University, [††]Microsoft Research Cambridge | [✉]maasch@cs.cornell.edu

Compositional Causal Reasoning

Overview Causal reasoning and compositional reasoning are two core aspirations in AI. Measuring the extent of these behaviors requires principled evaluation methods. A compositional view enables the systematic evaluation of causal reasoning in language models (LMs), revealing taxonomically distinct error patterns.

Def. 1 (Compositionality). When a measure f can be expressed as a function of measures $\{g_i\}_{i=1}^{n\geq 2}$.

Example 1. Decomposition of the total effect in linear models.



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

Def. 2 (Compositional causal reasoning (CCR)). The ability to infer (**A**) how local causal measures *compose* into global causal measures and (**B**) how global causal measures *decompose* into local causal measures, in both factual and counterfactual worlds.

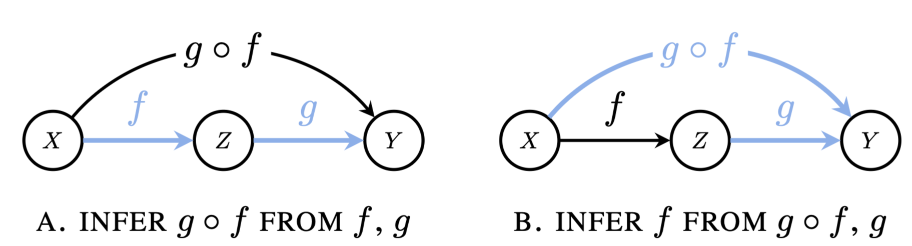


Figure 1: Commutative diagrams of inductive (**A**) and deductive (**B**) CCR.

Def. 3 (Compositional consistency). Reasoning is *compositionally consistent* when equivalent compositions are inferred to be equal.

Def. 4 (External validity). Reasoning is *externally valid* when estimates are equivalent to ground truth, up to error δ for metric θ:

$$\theta(\underbrace{\varphi^*_{\mathbf{x}}}_{\textit{true}},\ \underbrace{\widehat{\varphi}_{\mathbf{x}}}_{\textit{estimate}}) \leq \delta. \quad (1)$$

Def. 5 (Internal consistency). Reasoning is *internally consistent* when quantities that are theoretically equivalent are inferred to be equivalent, up to some error δ:

$$\underbrace{\varphi^*_{\mathbf{x}} = \varphi^*_{\mathbf{x}'}}_{\textit{equal in truth}} \;\Rightarrow\; \underbrace{\theta(\widehat{\varphi}_{\mathbf{x}}, \widehat{\varphi}_{\mathbf{x}'}) \leq \delta.}_{\sim\textit{equal in estimation}} \quad (2)$$

Def. 6 (Taxonomy of reasoners). Following from Definitions 4 and 5, we enumerate four distinct error patterns.

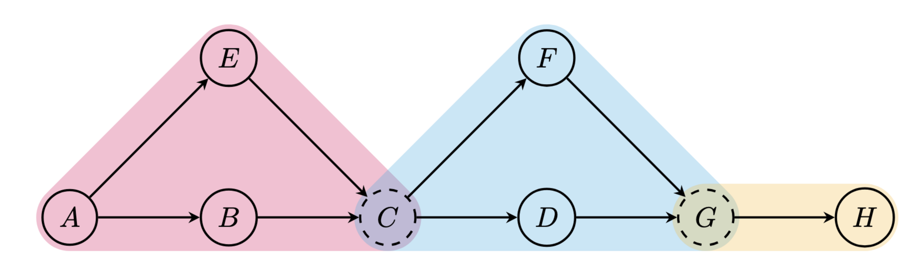
External validity \ *Internal consistency*		
	Valid-consistent (VC)	Valid-inconsistent (VI)
	Invalid-consistent (IC)	Invalid-inconsistent (II)

Demonstration: The PNS in Graphs with Cutpoints

Def. 7 (Probability of necessity and sufficiency (PNS), Pearl 1999). Let X and Y denote binary random variables, where X is a cause of Y. The probability that event x ($X = true$) is necessary and sufficient to produce event y ($Y = true$) is $\mathrm{PNS} \coloneqq \mathbb{P}(y_x, y'_{x'})$. When Y is *monotonic* in X, the PNS is identifiable as:

$$\mathrm{PNS} = \mathbb{P}(y_x) - \mathbb{P}(y_{x'}) = \mathbb{P}(y \mid do(x)) - \mathbb{P}(y \mid do(x')) = \mathrm{ATE}. \quad (3)$$

We exploit the following compositional property of the PNS.



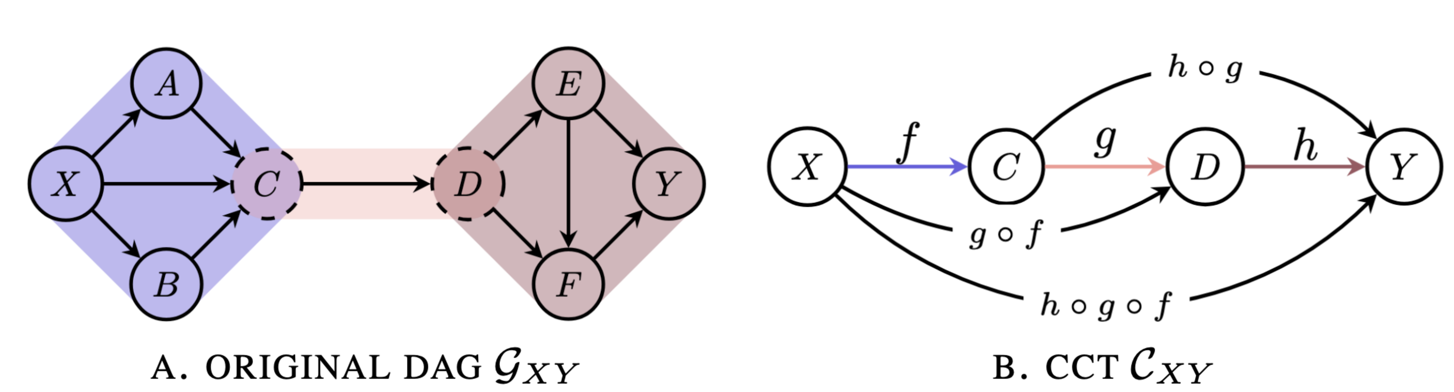
$$\mathrm{PNS}_{AH} = (\mathrm{PNS}_{AC} \cdot \mathrm{PNS}_{CG} \cdot \mathrm{PNS}_{GH}) = (\mathrm{PNS}_{AG} \cdot \mathrm{PNS}_{GH}) = (\mathrm{PNS}_{AC} \cdot \mathrm{PNS}_{CH})$$

Figure 2: Multiplicative composition over biconnected components (BCCs). Assume monotonicity.

For ease of exposition, we assume that the causal directed acyclic graph (DAG) has one root, one leaf, and no latent confounding.

Def. 8 (Commutative cut tree (CCT)). Given DAG $\mathcal{G}_{XY}$ with root X and leaf Y, CCT $\mathcal{C}_{XY}$ is obtained by a two-step transformation of $\mathcal{G}_{XY}$: **(1)** Construct a chain graph of the topological sort of X, the cutpoints in $\mathcal{G}_{XY}$, and Y; **(2)** Add edges to obtain a complete graph, and orient edges so all directed paths point from X to Y.

CCR as reasoning that CCTs commute: Given a causal measure that composes according to an associative function over BCCs, every composition corresponding to the paths from X to Y in $\mathcal{C}_{XY}$ should be equivalent to each other and to ground truth, up to some error.



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

Figure 3: (**A**) The DAG representing the running example used in our experiments. (**B**) Its corresponding CCT. (**Bottom**) The quantities of interest for CCR evaluation, derived from the CCT.

Results: Taxonomically Distinct Error Patterns

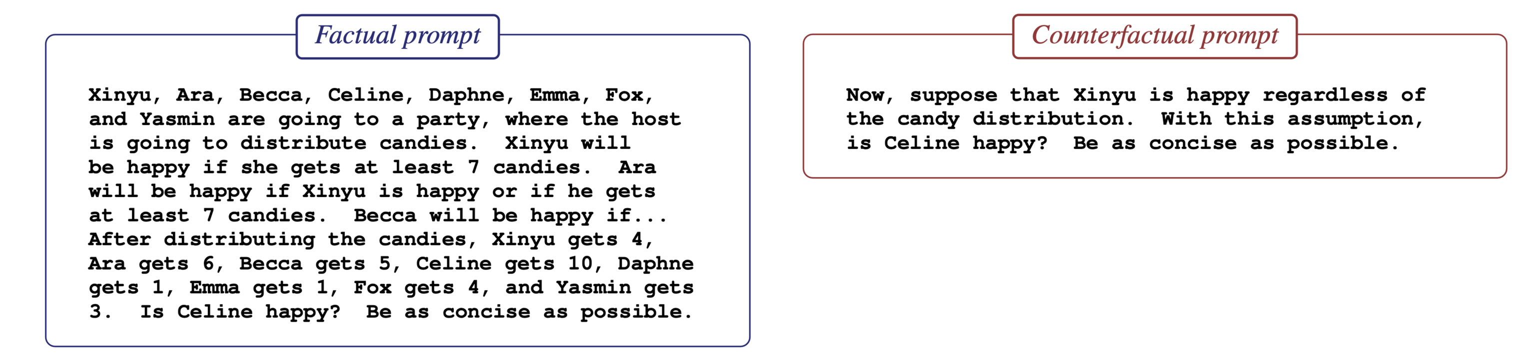


Figure 4: A `CandyParty` prompt whose narrative structure encodes the DAG in Figure 3A. To assess CCR at the counterfactual rung of Pearl's Causal Hierarchy, we treat LMs as *counterfactual data simulators*: instead of directly prompting the LM to perform formal causal inference, responses to series of factual and counterfactual "yes/no" questions were used to compute the PNS. We can obtain $\widehat{\mathrm{PNS}}_{XC}$ by simulating potential outcomes $X = true$, $X = false$ (Xinyu is or is not happy) and then querying for the value of C (Celine is or is not happy). Analogously, we obtain $\widehat{\mathrm{PNS}}_{DY}$ with interventions on D (Daphne's happiness) and queries on Y (Yasmin's happiness), etc.

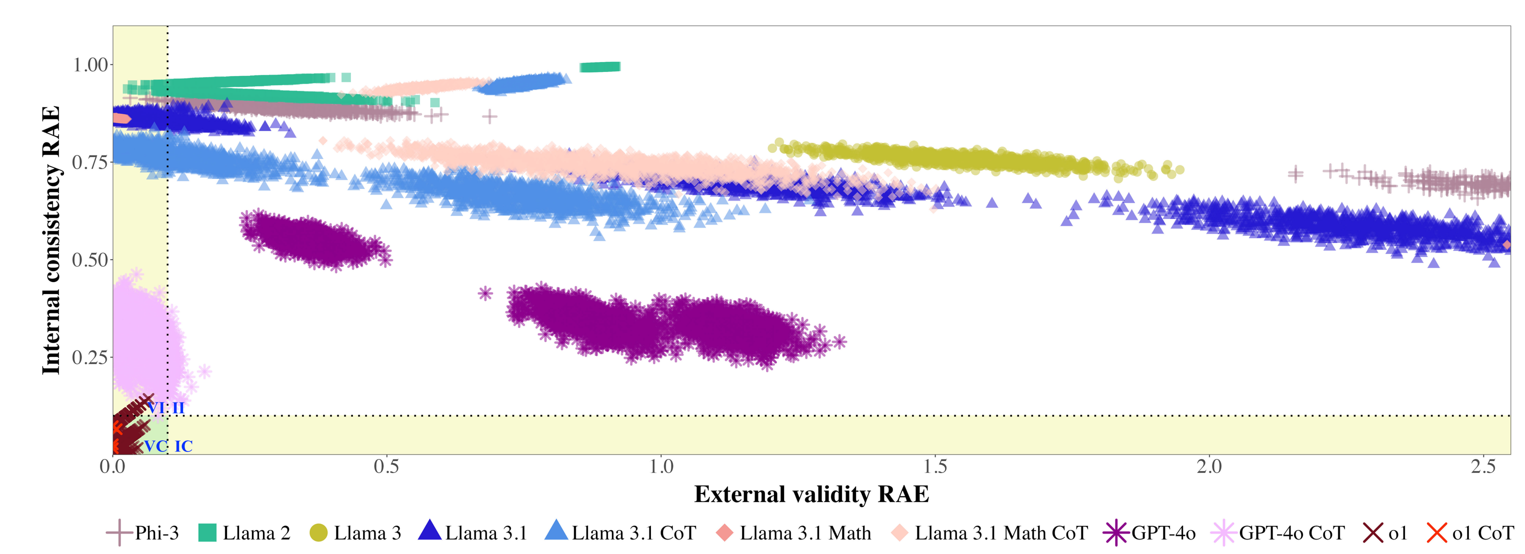


Figure 5: For PNS compositions ($n = 1000$ estimates per quantity per model), we compare relative absolute error (RAE) w.r.t. ground truth (external validity) and $\widehat{\mathrm{PNS}}_{XY}$ (internal consistency) to visualize our four reasoning quadrants (VI/IC in yellow; VC in green; II in white). Dotted lines are error thresholds (RAE = 0.1). Models are listed by increasing size. Note that the x-axis is truncated.

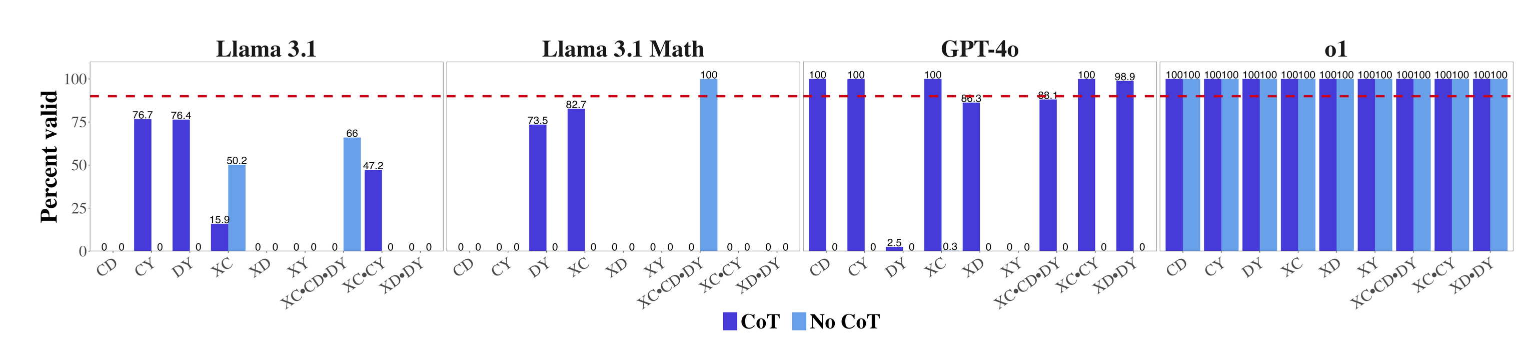


Figure 6: Percent of PNS estimates ($n = 1000$) that were externally valid for CoT vs non-CoT prompting. Reasoning was externally valid if $\geq 90\%$ of estimates had RAE ≤ 0.1 (red dashed line).

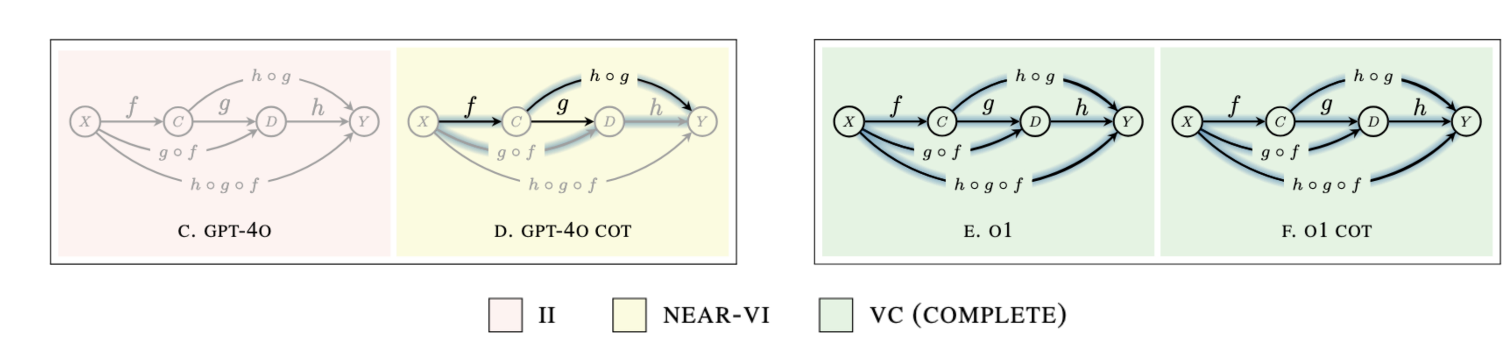


Figure 7: Visualizing (in)complete CCR, where only o1 reasons that $\mathcal{C}_{XY}$ commutes. Black edges are externally valid global and local quantities; gray are invalid. Paths from X to Y highlighted blue are externally valid compositions. Nodes are black when all paths passing through them are valid.

Presented at the 42[nd] ICML, Vancouver, Canada.