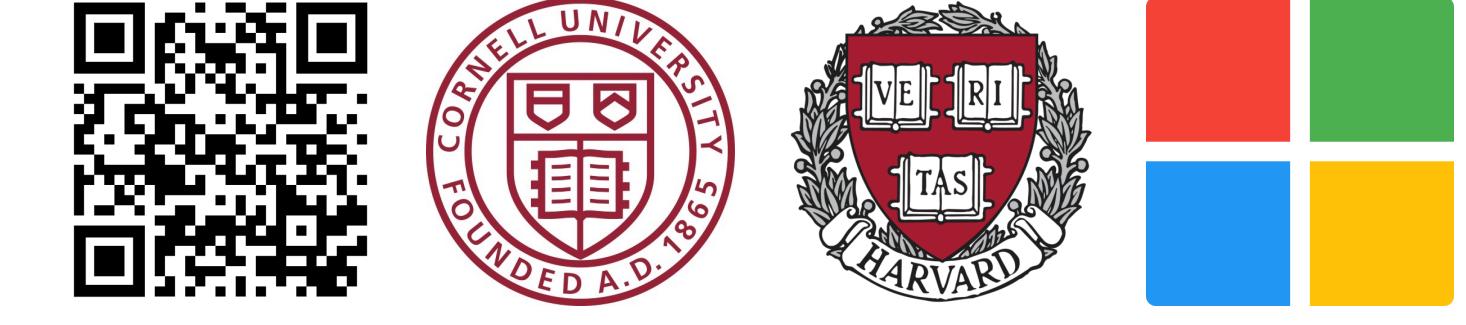


Compositional Causal Reasoning Evaluation in Language Models

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 (CCR)

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

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

Example 2. Decomposition of the total effect in linear models.

$$\text{TE}_{XY} = a + bc \quad \text{total effect} = \text{NDE}_{\text{direct effect}} + \text{NIE}_{\text{indirect effect}}$$

Definition 3 (Compositional causal reasoning (CCR)). The ability to correctly 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.

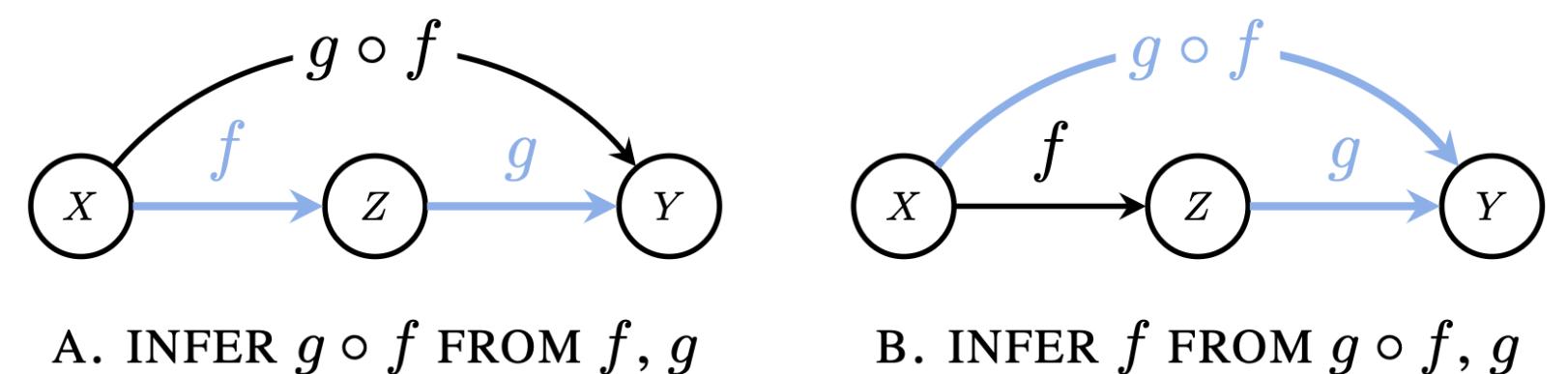


Figure 1: Commutative diagrams of inductive (A) and deductive (B) CCR.

Definition 4 (Compositional consistency). Theoretically equivalent compositions are inferred to be equal.

Definition 5 (External validity). Reasoning is *externally valid* when estimates are equivalent to ground truth, up to error δ for metric θ :

$$\theta(\varphi_x^*, \hat{\varphi}_x) \leq \delta. \quad (1)$$

true estimate

Definition 6 (Internal consistency). Reasoning is *internally consistent* when quantities that are theoretically equivalent are inferred to be equivalent, up to some error δ :

$$\varphi_x^* = \varphi_{x'}^* \Rightarrow \theta(\hat{\varphi}_x, \hat{\varphi}_{x'}) \leq \delta. \quad (2)$$

equal in truth equal in estimation

Valid-consistent (VC)	Valid-inconsistent (VI)
Invalid-consistent (IC)	Invalid-inconsistent (II)

Internal consistency

Figure 2: Taxonomy of reasoners: Following from Defs. 5, 6, we obtain four CCR error patterns.

Demonstration: The PNS in Graphs with Cutpoints

Definition 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 ($X = \text{true}$) is necessary and sufficient to produce event y ($Y = \text{true}$) is $\text{PNS} := \mathbb{P}(y_x, y'_x)$. When Y is *monotonic* in X , the PNS is identifiable as:

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

We exploit the following compositional property of the PNS.

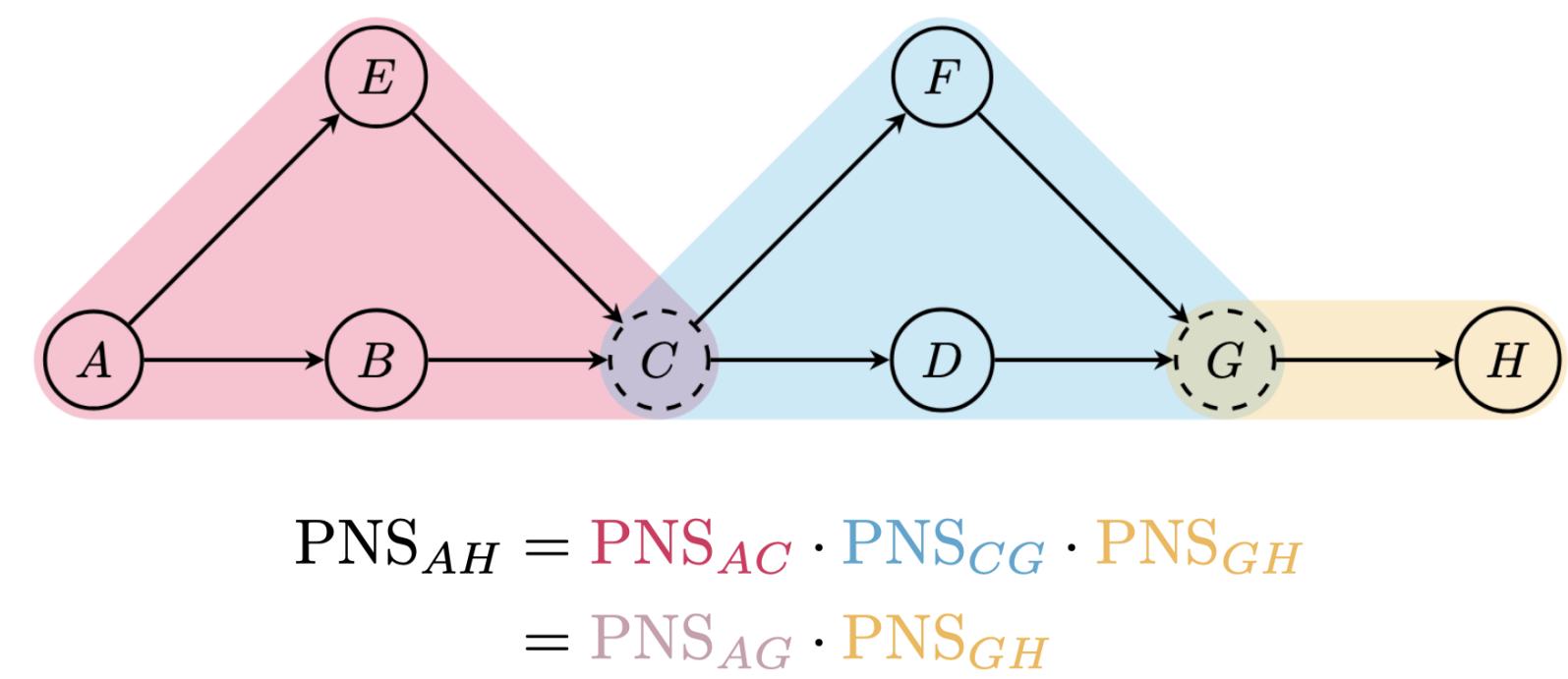


Figure 3: Multiplicative composition over biconnected components (BCCs). Assume monotonicity.

For ease of exposition, we assume problem structures whose causal DAGs have one root, one leaf, and no latent confounding.

Definition 8 (Commutative cut tree (CCT)). Given DAG \mathcal{G}_{XY} and a causal measure that composes according to an associative function over BCCs, 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 the root, cutpoints, and leaf in \mathcal{G}_{XY} ; (2) Add a directed edge between non-adjacent nodes in the chain, yielding a complete graph where all directed paths point from X to Y .

CCR as reasoning that \mathcal{C}_{XY} commutes: 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.

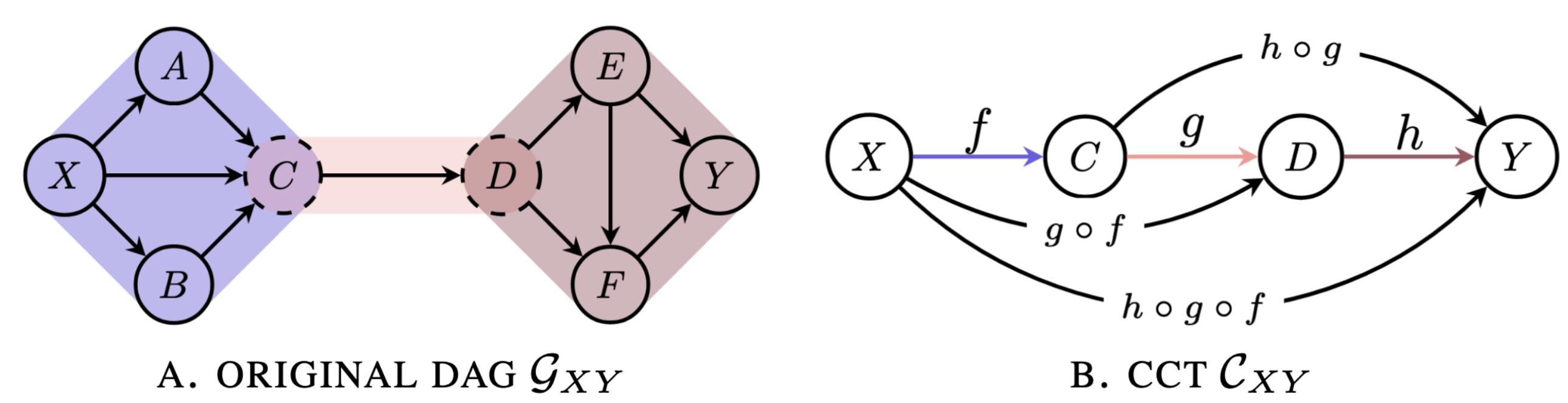


Figure 4: (A) The DAG representing the running example used in our experiments. (B) Its corresponding CCT. (Bottom) The quantities of interest for CCR evaluation, derived from the CCT.

Results: CCR Evaluation in Seven LMs

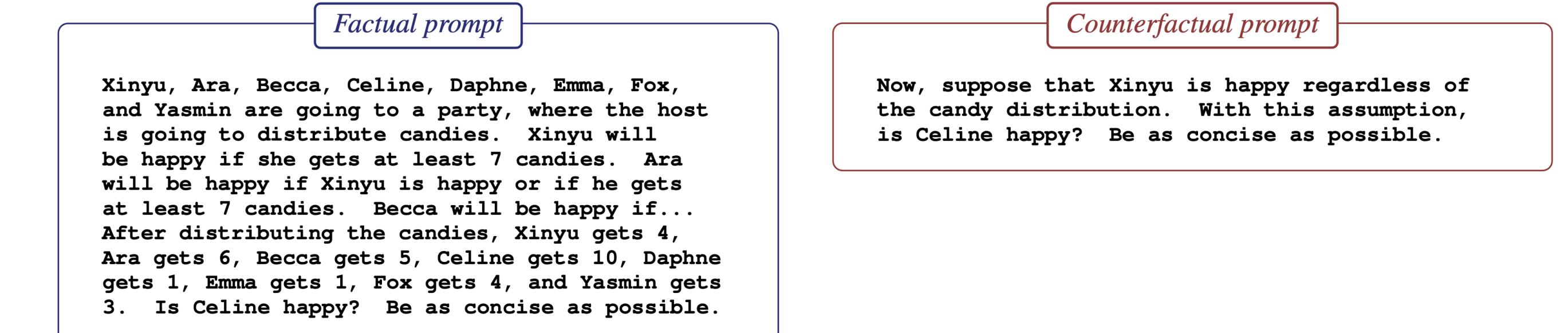


Figure 5: A CandyParty prompt whose narrative structure encodes the DAG in Figure 4A. To assess CCR at the counterfactual rung of Pearl's Causal Hierarchy, we treat LMs as *counterfactual data simulators*: instead of directly prompting the LM to perform formal causal inference, responses to series of factual and counterfactual “yes/no” questions were used to compute the PNS. We can obtain 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 PNS_{DY} with interventions on D (Daphne’s happiness) and queries on Y (Yasmin’s happiness), etc.

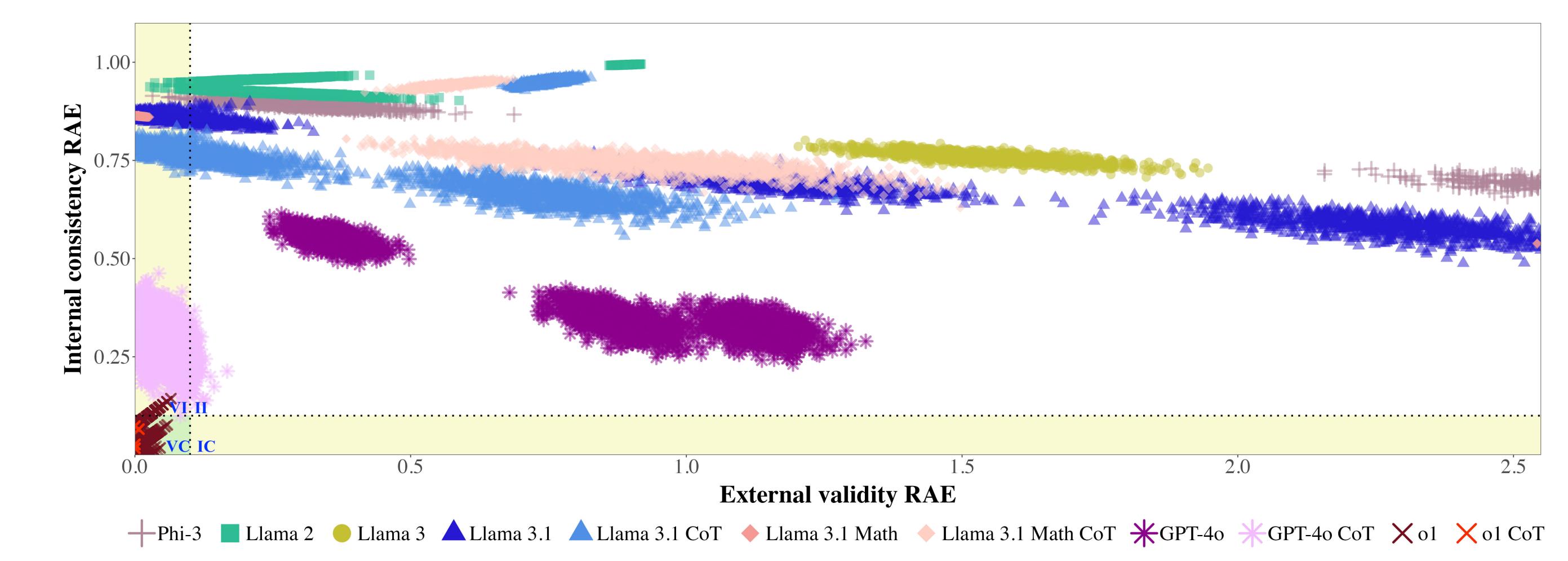


Figure 6: For PNS compositions ($n = 1000$ estimates per quantity per model), we compare relative absolute error (RAE) w.r.t. ground truth (external validity) and PNS_{XY} (internal consistency) to visualize our four reasoning quadrants (VI/IC in yellow; VC in green; II in white). Dotted lines are error thresholds (RAE = 0.1). Models are listed by increasing size.

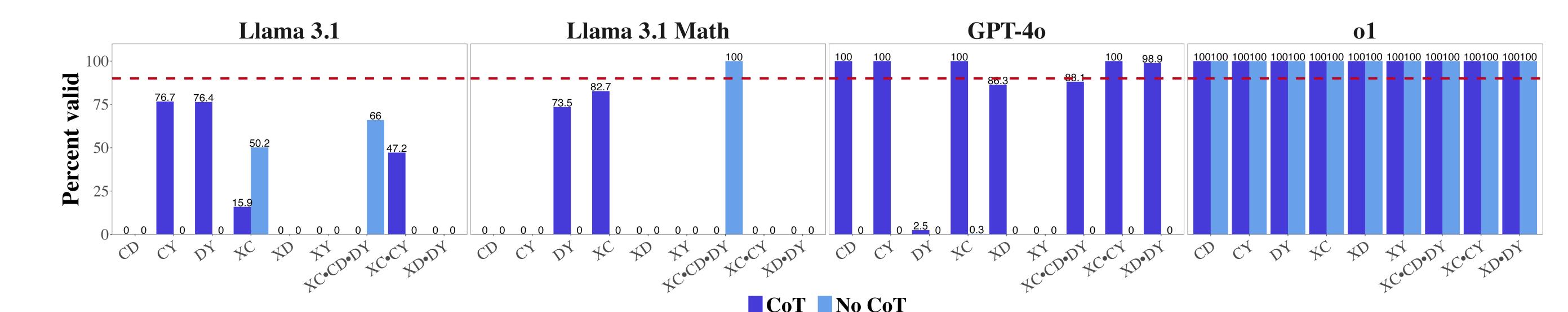


Figure 7: Percent of PNS estimates ($n = 1000$) that were externally valid for CoT vs non-CoT prompting. Reasoning was externally valid if $\geq 90\%$ of estimates had RAE ≤ 0.1 (red dashed line).

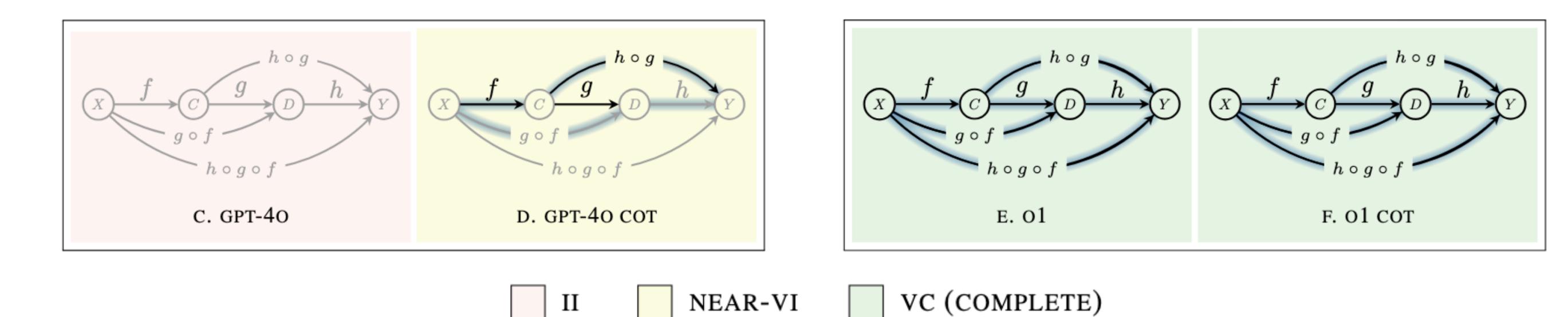


Figure 8: Visualizing (in)complete CCR, where only o1 reasons that \mathcal{C}_{XY} commutes. Black edges are externally valid global and local quantities; gray are invalid. Paths from X to Y highlighted blue are externally valid compositions. Nodes are black when all paths passing through them are valid.