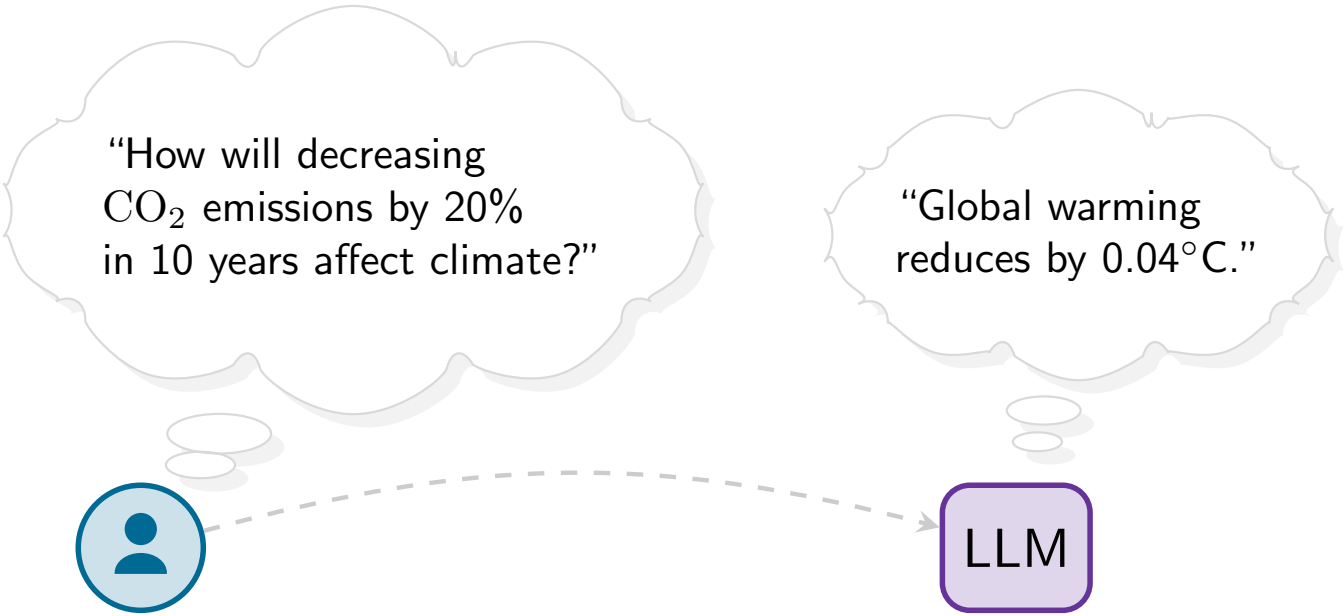
Evaluating Bivariate Causal Statements based on Mutual Compatibility

Erik Jahn[1, 2] Dominik Janzing[2]

[1]Amazon Research Tübingen, Germany
[2]California Institute of Technology, USA

Motivation



Question

How can we falsify causal claims without experiments?

Idea

If we have many causal statements, we can try to check their compatibility!

Setup

We study **complete acyclic** lists of **linear bivariate** causal statements for n variables:

- **Bivariate:** Each statement is on a pair of variables (X, Y).
- **Linear:** Each statement specifies a causal direction $X \to Y$ and a linear causal coefficient α: If X is shifted by δ, then Y shifts by $\alpha\delta$.
- **Complete:** We have a causal statement for each of the $\binom{n}{2}$ possible pairs of variables.
- **Acyclic:** The causal directions are consistent with a total ordering of the variables.

We also have **observational data** for the n variables.

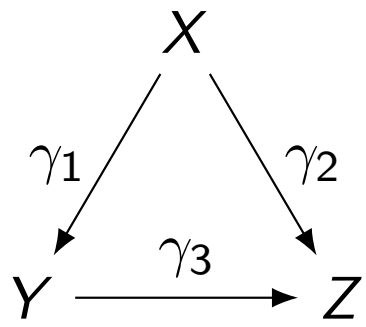
Induced Multivariate Model

Lemma

Each complete acyclic list of linear bivariate causal statements induces a unique multivariate linear causal model that fits the data and whose pairwise marginals coincide with the bivariate statements.

$$X \xrightarrow{\alpha_1} Y$$
$$X \xrightarrow{\alpha_2} Z$$
$$Y \xrightarrow{\alpha_3} Z$$

$\Longrightarrow$

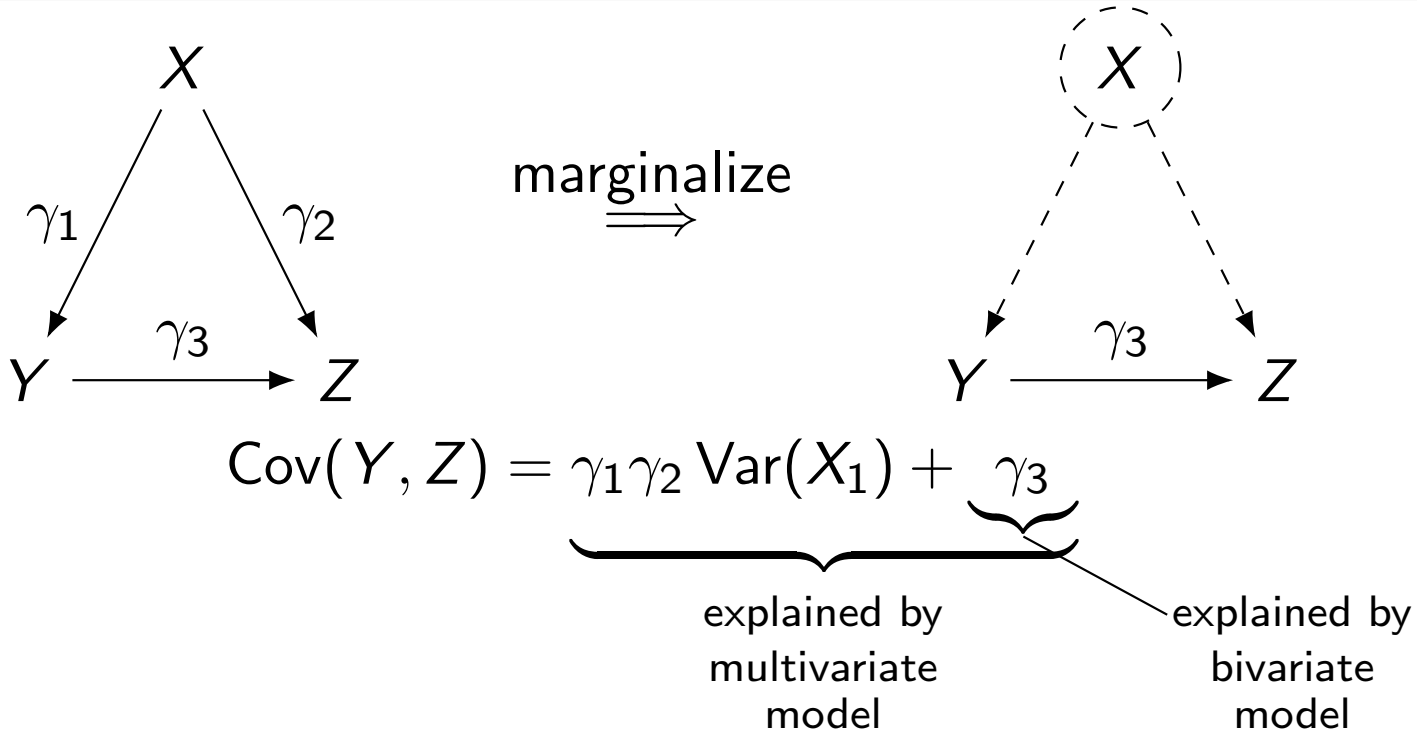


$$\alpha_1 = \gamma_1$$
$$\alpha_2 = \gamma_2 + \gamma_1\gamma_3$$
$$\alpha_3 = \gamma_3$$

Model Plausibility

Postulate

A plausible multivariate model should be able to explain pairwise covariances better than any individual bivariate model by itself.



Compatibility Score

For a list A of linear causal statements, we define:

- $\mathcal{C}^{\text{biv}}(X_i, X_j, A)$ = squared part of $\text{Cov}(X_i, X_j)$ that is unexplained by the bivariate model
- $\mathcal{C}^{\text{mult}}(X_i, X_j, A)$ = squared part of $\text{Cov}(X_i, X_j)$ that is unexplained by the multivariate model

Theorem

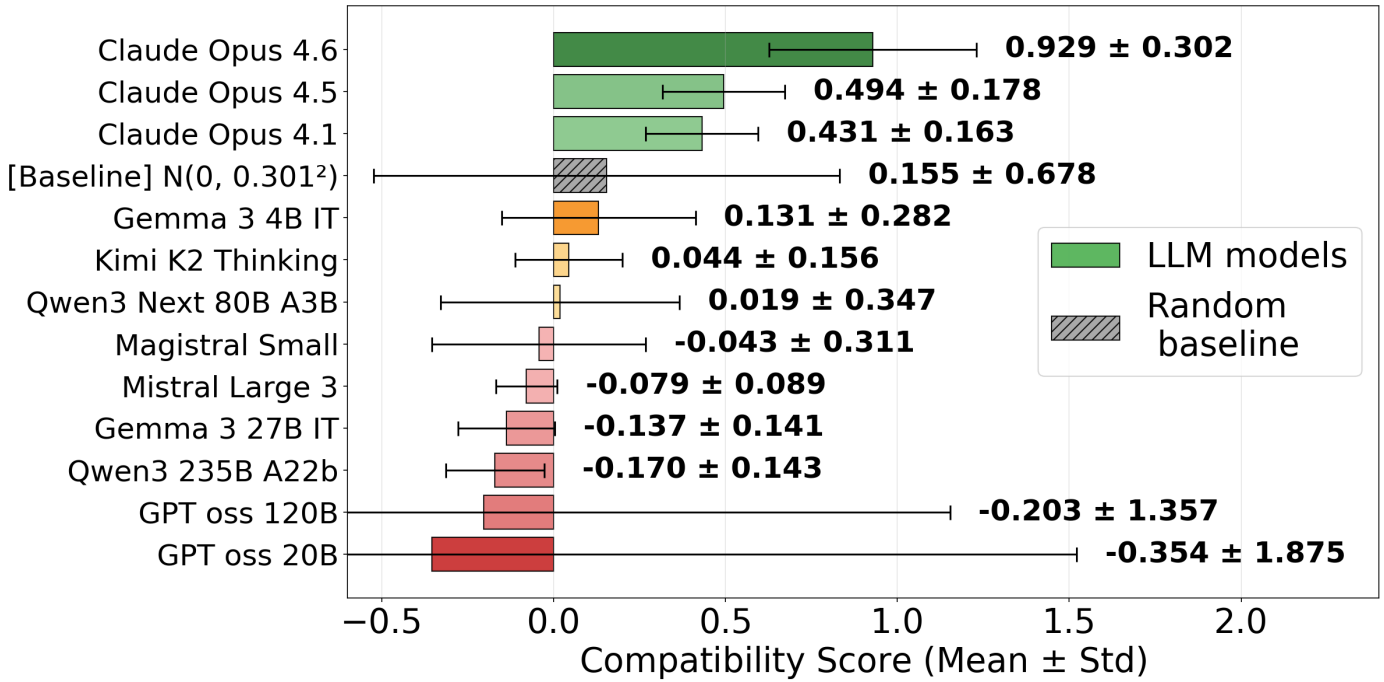
For random causal models we have $\sum_{i,j} \mathcal{C}^{biv}(X_i, X_j, A^*) > \sum_{i,j} \mathcal{C}^{mult}(X_i, X_j, A^*)$ for the true causal statements A^* in expectation (under an independence of mechanisms assumption).

Definition

We define the compatibility score for linear bivariate causal statements as

$$\text{comp}(X, A) := \sum_{i<j} \left(\mathcal{C}^{biv}(X_i, X_j, A) - \mathcal{C}^{mult}(X_i, X_j, A) \right).$$

Compatibility Scores of LLMs



Causal statements generated given data and variables from the Gapminder Foundation(`https://www.gapminder.org/data/`):

- population density,
- literacy rate,
- average daily income,
- access to basic sanitation,
- percentage of adults that smoke
- happiness score,
- average life expectancy.

Check out the paper!

https://arxiv.org/abs/2606.00278