Causal Inference and Large Language Models

STAT9911

Kai Cooper

April 24, 2025

Operations, Information and Decisions

HEALTH & WELLNESS

A boy saw 17 doctors over 3 years for chronic pain. ChatGPT found the diagnosis

Alex experienced pain that stopped him from playing with other children but doctors had no answers to why. His frustrated mom asked ChatGPT for help.

Figure 1: Dr. GPT

- Did ChatGPT infer causality from this boy's medical data?

Brief Primer on Causal Inference

Causality

- Causal inference is primarily concerned with the determining the effects of known causes.

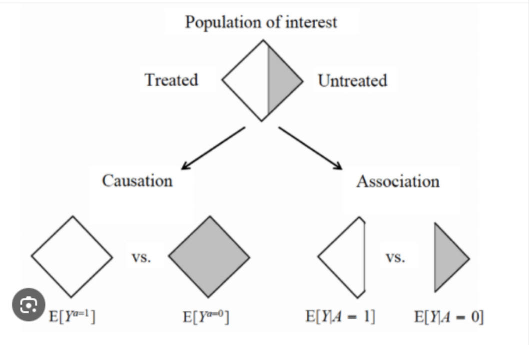


Figure 2: Causation versus association

SCMs

Definition: (Pearl, 2009) A Structural Causal Model (SCM) is a tuple $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{F}, P(\mathcal{U}))$ where:

- $\mathcal{U}$: set of exogenous variables (relevant "nature")
- $\mathcal{V}$: set of endogenous variables (relevant but measurable)
- $\mathcal{F}$: set of functions $\{V_i = f_i(\mathrm{pa}_i, U_i)\}$, one for each $V_i \in \mathcal{V}$
- $P(\mathcal{U})$: probability distribution over exogenous variables

Example: Consider variables $\mathcal{V} = (X, Y, Z)$.

- Structural equations:

$$X = f_X(U_X) \quad Y = f_Y(X, U_Y) \quad Z = f_Z(Y, U_Z)$$

- DAG associated with the causal model:

$$X \longrightarrow Y \longrightarrow Z$$

Potential outcomes

Definition (Potential Outcome)
Given a structural causal model $\mathcal{M}$, the potential outcome $Y(x)$ is defined as the value of Y obtained by intervening to set $X = x$ and recursively evaluating the functions in $\mathcal{F}$:

$$Y(x) := Y_{\mathcal{M}[X \leftarrow x]}(U)$$

where $\mathcal{M}[X \leftarrow x]$ denotes the modified SCM in which the function f_X is replaced by the constant x.

Definition (Individual Causal effect)
Suppose we have a set of individuals under study indexed by i. The individual causal effect is a comparison

$$Y_i(x) \text{ vs } Y_i(x')$$

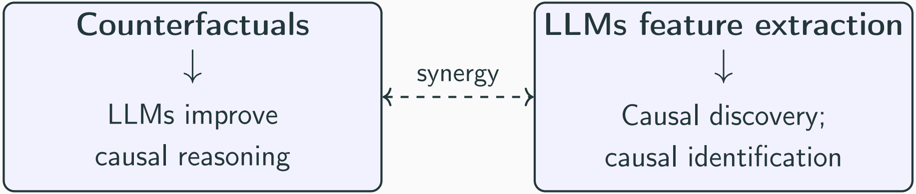
where $x \neq x'$.

The fundamental problem isn't so problematic

- *Fundamental problem of causal inference* (Holland, 1986) is that we cannot ever compute the individual level causal effect...
- ...**for people**, e.g. I cannot both give you a drug and its placebo.
    - Often get sidestep this issue by randomisation, which identifies population-level effects instead.
- For machines whose architecture we understand well, this is less of an issue.
- Old adage: "no causation without manipulation". Take this to a new level.

Causality and LLMs overview

- LLMs excel at pattern recognition and knowledge retrieval, but causal reasoning is known to be difficult for them.
- However, this creates an intriguing synergy between causality and LLMs.



Outline

- **Understanding LLMs with causal inference**: using causal methods to interpret LLM behavior
- **Improving LLMs with causal inference**: using ideas from causal inference to improve causal reasoning and prevent bias in LLMs.
- **Using LLMs for causal inference**: employing LLMs as tools for causal identification strategies
- **Conclusion**.

Understanding LLMs with Causal Inference

Interpreting Model Output

- Traditional black-box model interpretation often finds correlations (e.g. probing hidden states for contained information) but cannot confirm if a component *causes* a behavior.
- Many approaches to interpretation have been described as *causally-aware*
  - Think: Shapley values; partial dependence plots; etc.
- Appeal to causal inference to improve "explainability"
- One approach—**causal mediation analysis**: Treat internal components of the model as mediators in a causal graph connecting inputs to outputs. By intervening on these mediators, we can measure:
  - **Direct effect**: Influence of input on output not through the mediator.
  - **Indirect effect**: Influence transmitted via the mediator.
- This helps identify which neurons or attention heads are *causally responsible* for certain model outputs, not just correlated with them.

Causal Mediation in Neural LMs

Setup: We model the causal pathway from input X to output Y via mediator Z (e.g., neuron, attention head).

$$
\begin{tikzcd}
X \arrow[r] \arrow[rr, bend left] & Z \arrow[r] & Y
\end{tikzcd}
$$

Decomposition of Treatment Effect:

(What?)

- **Total Effect (TE):** $\mathbb{E}\left[Y(X=1) - Y(X=0)\right]$

(Why?)

- **Natural Direct Effect (NDE):** $\mathbb{E}\left[Y(X=1, Z(X=0)) - Y(X=0, Z(X=0))\right]$
- **Natural Indirect Effect (NIE):** $\mathbb{E}\left[Y(X=1, Z(X=1)) - Y(X=1, Z(X=0))\right]$

Case Study: Gender Bias in GPT-2 (Vig et al., 2020)

- Hypothesis: Language models often prefer gender-stereotypical pronouns in ambiguous contexts (e.g. for "The nurse said that...", the model is likelier to fill "she" than "he").
- Test: adapted causal mediation analysis to a Transformer model. They treated the presence of gender information in the input as the "treatment" and analyzed internal components (neurons/attention heads) as mediators carrying gender bias.

Experiment

Key Idea: Treat individual model components (e.g., neurons, attention heads) as mediators Z.

- Forward pass: $X \to H \to Y$, where H includes layers/heads/neurons.
- Define $Z \subset H$ as the targeted mediator (e.g., a single head).
- Intervention on Z is simulated by overwriting Z with counterfactual values.

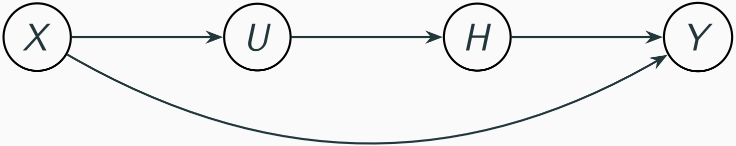
Procedure:

1. Run the model on $X = x_0$ to extract baseline activations $Z(X = 0)$.
2. Run the model on $X = x_1$, then override $Z \leftarrow Z(X = 0)$ to compute $Y(X = 1, Z(X = 0))$.
3. Likewise compute $Y(X = 0, Z(X = 1))$.

This manipulation enables computation of NDE and NIE. Note that these quantities are impossible to get in a standard RCT.

Treatment X is the focal word, e.g. “doctor" or “man". The prompt U is the fixed prefix containing this word, e.g. "The x said that...". Y is next token.

DAG:



Target

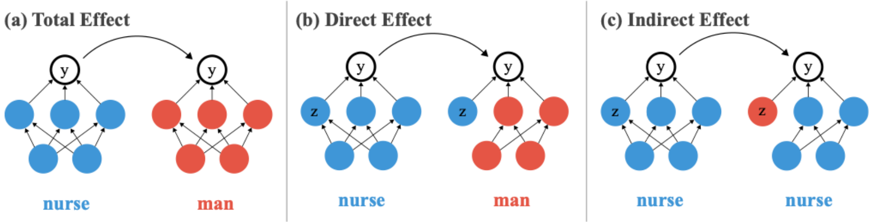


Figure 2: Mediation analysis illustration. Here the *do*-operation is $x = \texttt{set-gender}$, which changes u from *nurse* to *man* in this example. The **total effect** measures the change in y resulting from the intervention; the **direct effect** measures the change in y resulting from performing the intervention while holding a mediator z fixed; the **indirect effect** measures the change caused by setting z to its value under the intervention, while holding u fixed.

Figure 3:

- First experiment $Z \subset H$ are neurons; second experiment Z are attention weights.

Target

The targets of measurement are

$$\eta_e = \frac{\Pr(C = \text{anti-stereotypical})}{\Pr(C = \text{stereotypical})} \Big/ \frac{\Pr(C' = \text{anti-stereotypical})}{\Pr(C' = \text{stereotypical})} - 1$$

where $e \in \{\mathrm{TE}, \mathrm{NDE}, \mathrm{NIE}\}$ and C and C' are the appropriate counterfactuals for comparison. For example, for the NIE,
$C = Y(X = \text{doctor}, Z = (\text{neuron under } X = \text{man}))$ and $C' = Y(X = \text{doctor})$.

Example

u = The nurse said that [blank]

1) Compute relative probabilities of the baseline.

$p([he]|u) = p([he]|\text{the nurse said that}) \approx 0.03$

$p([she]|u) = p([she]|\text{the nurse said that}) \approx 0.22$

$\boldsymbol{y}_{\texttt{null}}(u) = 0.03/0.22 \approx 0.14$

3) Compute the total effect

$\mathrm{TE}(\texttt{set-gender}, \texttt{null}; \boldsymbol{y}, u)$
$= 13.1/0.14 - 1 \approx 92.6$

2) Set u to an anti-stereotypical case and recompute.

$x = \texttt{set-gender}$: change nurse $\rightarrow$ man

$p([he]|u, \texttt{set-gender}) =$
$p([he]|\text{the man said that}) \approx 0.32$

$p([she]|u, \texttt{set-gender}) =$
$p([she]|\text{the man said that}) \approx 0.02$

$\boldsymbol{y}_{\texttt{set-gender}}(u) = 0.32/0.02 \approx 13.1$

Data and setup

Models: GPT2 variants (small to XL)

Datasets:

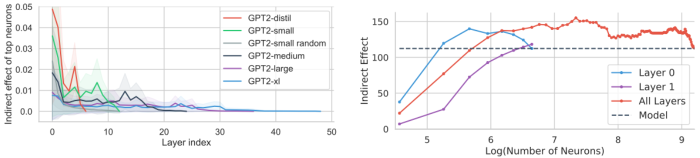
- **Professions**: prompts with stereotyped occupations
- **WinoBias, WinoGender**: Winograd-style coreference sentences, e.g.

Prompt u: The nurse examined the farmer for injuries because she ____
Stereotypical candidate: was caring
Anti-stereotypical candidate: was screaming

Figure 5: WinoGender example

Key Findings:

- **Sparsity:** Gender bias is mediated by a small set of neurons/heads
- **Localization:** Neurons in early layers, heads in mid-layers (e.g., layer 4–5)
- **Scale effect:** Larger models (e.g., GPT2-XL) exhibit stronger bias
- Bias is traceable and localized—enabling mitigation via targeting specific components.



(a) Indirect effects of top neurons in different models on the professions dataset. Here, early layers have the largest effect.

(b) Indirect effects after sequentially selecting an increasing number of neurons from either the full model or individual layers using the TOP-K approach in GPT2-small on the professions dataset.

Figure 7: Sparsity effects in neurons.

Case Study: Robustness in Math Reasoning (Stolfo et al., 2023)

- Hypothesis: Do larger LLMs actually reason better, or do they just pick up on superficial cues? Stolfo et al. (2023) examine robustness of LLMs on math word problems.
- Test: Construct a causal graph of factors influencing an LLM's success on a math problem:
  - Surface form of the problem (linguistic style, irrelevant details).
  - The specific operands/numbers used.
  - The operation or reasoning steps required (e.g. addition vs multiplication).
  - Y: Whether the model arrives at the correct solution.
- By intervening on each factor (e.g., rephrasing the problem without changing the math, or changing the numbers while keeping structure), they measure the causal effect of that factor on Y.
- A model is *robust* if Y doesn't change when surface form is altered, but is appropriately sensitive when relevant factors change.

Case Study: Robustness in Math Reasoning (Stolfo et al., 2023)

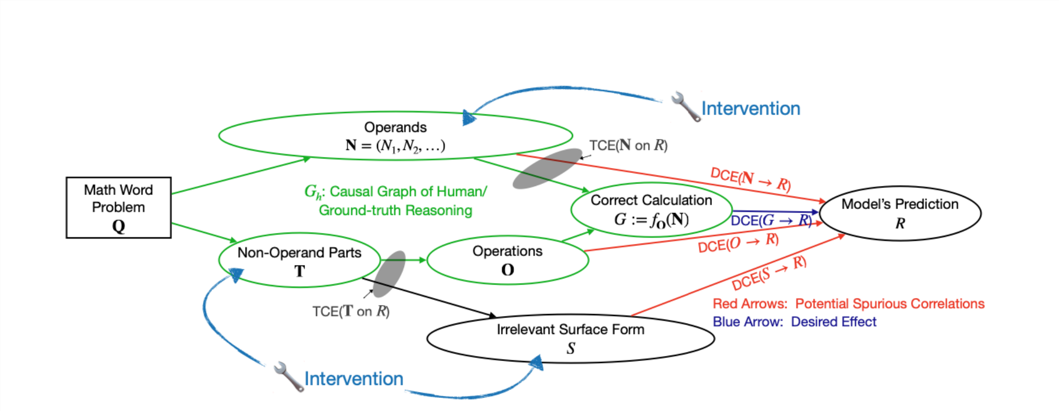


Figure 7: DGP

Findings

- Smaller models often latch onto spurious cues:
  - E.g. rephrasing a problem or adding irrelevant text would often alter their answer, indicating over-reliance on surface form.
- As model size increases, robustness does not steadily improve at first:
  - Some larger models still showed sensitivity to superficial changes.
  - Simply scaling up from, say, 6B to 13B parameters did not guarantee significantly better causal robustness.
- Dramatic jump at GPT-3 (175B): The 175B GPT-3 (especially the instruction-tuned version) was markedly more robust:
  - It was largely invariant to rewording or distractions.
  - It remained appropriately sensitive to the actual math content, solving problems by focusing on the right factors.
- "Emergent abilities hypothesis".

Improving LLMs with Causal Inference

Chain of thought reasoning

Goal: Improve the **faithfulness** and **causal consistency** of foundation models (FMs) in knowledge reasoning.

Challenges:

- Existing chain-of-thought (CoT) prompting leads to *inferential fallacies* and *hallucinations*
- No explicit modeling of causal dependencies in reasoning
- Lack of collaborative mechanisms to validate causal consistency

Proposal: CausalGPT with **CaCo-CoT** (Causal-Consistency Chain-of-Thought) framework (Tang et al., 2025)

- Multi-agent framework with faithful reasoners and causal evaluators
- Combines causal reasoning and collaborative validation

Overview

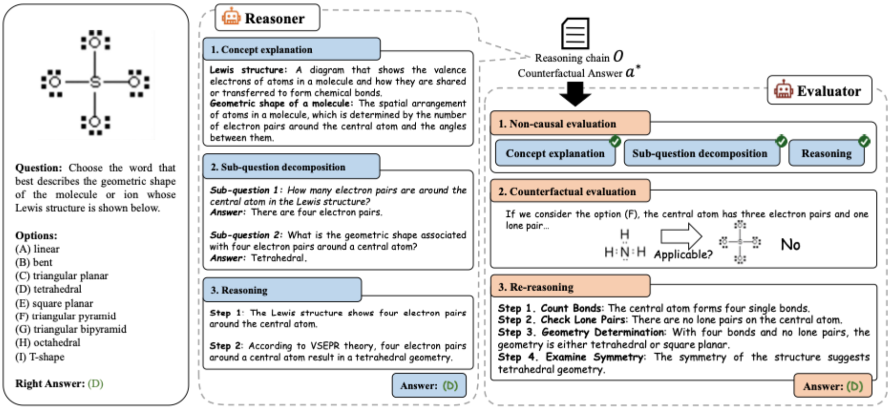


Fig. 3. Demonstration of how a *faithful reasoner* and a *causal evaluator* analyze a molecular geometry question.

Figure 8: Example

Faithful Reasoner Design

Mimic human-reasoning processes

1. **Concept Explanation:**
$$r_1 := \texttt{Expl}(R, Q)$$
2. **Sub-question Decomposition:**
$$r_2 := (q_1, \ldots, q_t) = \texttt{Decomp}(R, Q, r_1)$$
$q_t =$ "sub-question".
3. **Sub-question Answering:**
$$r_3 := \{e_t = \texttt{Solve}(R, q_t, r_1, r_2, e_{<t})\}$$
$e_t =$ "evidence"
4. **Final Answer:** $a_R = \texttt{Solve}(R, Q, r_1, r_2, r_3)$

Output: Reasoning chain $O = \{r_1, r_2, r_3, a_R\}$

Causal Evaluator

Goal: Check if the reasoning chain $O = \{r_1, r_2, r_3, a_R\}$ contains any factual or inferential errors.

Stepwise evaluation:

$$g_i := \texttt{Eval}(E, Q, O, g_{<i}) \in 0, 1 \quad \text{for } i = 1, 2, 3$$

- $g_i = 1$ if step r_i is valid given prior steps
- $g_i = 0$ if any inconsistency or error is found

Aggregate Score:

$$S_{\text{non}} = \frac{1}{3} \sum_{i=1}^{3} g_i$$

Purpose: Identify flawed reasoning not evident from local context alone by re-evaluating the full chain holistically. Called “non-causal" because it evaluates past with the future.

Causal Evaluator: Counterfactual Evaluation

1. **Construct a counterfactual premise**
   - Choose alternative answer $a^* \neq a_R$
   - Create counterfactual context: $c^* = \{a_R, a^*\}$
2. **Re-label evidence** $e = \{e_1, ..., e_T\}$ **with**

$$b_j \in \{-1, 0, 1\} \quad \text{for each } e_j$$

   - $b_j = 1$: supports both a_R and a
   - $b_j = 0$: contradicts under c (i.e., $e_j \wedge e_k \Rightarrow$ contradiction)
   - $b_j = -1$: irrelevant

Filter

$$e^* = \{e_j : b_j = 1\}$$

Causal Evaluator: Counterfactual Evaluation

3. **Solve using filtered evidence**

$$a_E^* := \texttt{Solve}(E, Q, c^*, e^*)$$

4. **Score the answer**

$$S_{\text{counter}} = \begin{cases} 1 & \text{if } a_E^* = a_R \\ 0 & \text{otherwise} \end{cases}$$

5. **Acceptance policy**

$$a_E = \begin{cases} a_R & \text{if } S_{\text{non}} + S_{\text{counter}} = 2 \\ \texttt{Solve}(E, Q, g, e^*) & \text{o/w} \end{cases}$$

Example

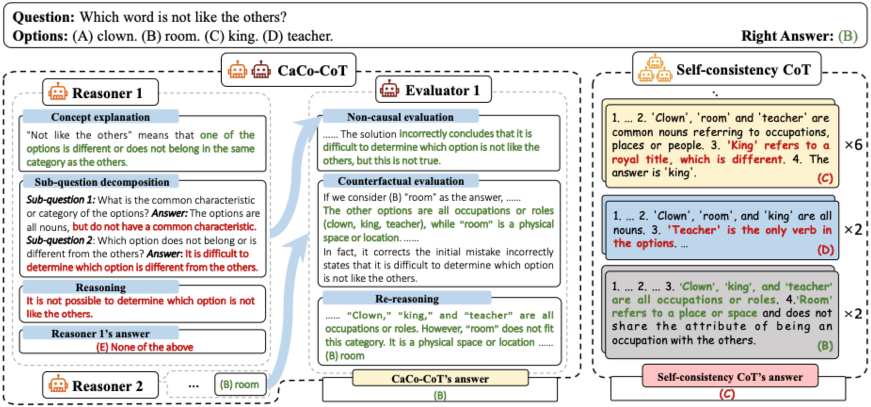


Fig. 6. Zero-shot CaCo-CoT versus zero-shot SC-CoT in resolving a question on ScienceQA. Text highlighted in green indicates positive contents that contribute to the correct answer, while the red portions represent unfaithful or misleading content.

Figure 9: Example

Extend with a multiple agents

Procedure:

1. Generate n reasoning chains $O_1, ..., O_n$
2. Majority voting yields top-1 answer a_{top1}
3. Evaluation:
    - If $\mathrm{P}(a_{\text{top1}}) > h_0$, evaluate one chain O_{top1}
    - If consistent, accept answer; else, recurse
    - If no consensus, evaluate top-k candidates with k evaluators

Empirical Performance

TABLE II

COMPARISON OF DIFFERENT COOPERATION MODES ON THE TEXTUAL SPLIT OF SCIENCEQA. NAT, SOC, AND LAN REFER TO NATURAL SCIENCE, LANGUAGE SCIENCE, AND SOCIAL SCIENCE RESPECTIVELY. THE LEVEL OF GRADE REPRESENTS THE DIFFICULTY OF THE QUESTION. † DENOTES ONE-SHOT SETTINGS, WHILE ITS ABSENCE INDICATES ZERO-SHOT SETTINGS. k IS THE MAXIMUM NUMBER OF *evaluators* TO BE INVOKED IN EACH ROUND.

Method	Category	Subject			Grade		Average
		NAT	SOC	LAN	G1-6	G7-12	
CoT [56]	Single-chain	85.71	86.40	77.37	82.70	80.52	81.79
CoT† [6]		85.91	92.00	82.95	86.41	82.67	84.85
L2M-CoT [33]		83.41	91.20	83.89	86.17	81.16	84.08
CaCo-CoT$_{R1}$ **w/o** *evaluator*		85.52	95.20	82.01	86.02	82.13	84.40
CaCo-CoT†$_{R1}$ **w/o** *evaluator*		89.26	96.00	81.44	88.34	82.56	85.93
SoT [69]	Ensembling	64.72	73.60	69.41	70.73	62.86	67.45
SC-CoT† [9]	Majority voting	88.78	91.20	83.33	88.57	83.21	86.33
C-CoT† [10]		88.78	91.20	84.47	89.03	83.85	86.87
CaCo-CoT$_{R3}$ **w/o** *evaluator*		87.44	92.00	83.43	87.49	83.42	85.79
CaCo-CoT†$_{R3}$ **w/o** *evaluator*		90.79	**99.20**	83.32	89.79	84.82	87.71
CaCo-CoT$_{R2}$ ($p = 2$)	Multi-round cooperation	88.78	93.60	83.43	88.57	83.64	86.51
CaCo-CoT†$_{R2}$ ($p = 2$)		91.37	97.60	**84.19**	90.12	85.79	88.31
CaCo-CoT$_{R3}$ ($p = 2$)		89.45	96.00	83.52	89.03	84.18	87.01
CaCo-CoT†$_{R3}$ ($p = 2$)		**93.38**	96.80	83.62	**90.97**	**86.11**	**88.94**

Figure 10: Performance

Future Work: Handle unfaithful consensus, improve efficiency, explore open-domain reasoning, mix-and-match language models

Improving LLMs via Causal Inference

Related to a paper examining the utility of causal inference for improving NLP methods. Ideas raised focus on integrating causal methods into the training of the models:

- counterfactual invariance: predictions should be invariant to irrelevant perturbations
- counterfactual fairness: protected class attributes should have zero causal effect on predictions

How to implement:

- Add as penalties in loss function
- Data augmentation, balance training data to reduce downstream causal effect

Caution: Sani et al. (2025), confounding by hidden true label.

Using LLMs for Causal Inference

Causal discovery

Goal: Learn the underlying causal structure (typically a DAG) from observational or interventional data.

- Data is assumed to be generated from a SCM
- The joint distribution $P(V)$ is Markov and faithful to the underlying DAG $\mathcal{G}$.

Challenge:

- Multiple DAGs can encode the same set of conditional independencies
- Without interventions or additional assumptions, we cannot uniquely identify the true causal graph.
    - e.g. one cannot tell apart $A \rightarrow B$ from $B \rightarrow A$ with $P(A, B)$ alone.
- Many algorithms exist, major categories: constraint (conditional independence tests), score based (e.g. AIC).

How to solve a causal problem

Recall: causal inference is about finding effects from interventions

- Step 1 is always to define the DGP.
- How does one do that? Expert knowledge, **conditional on recent history** (causal frame)
- Question: rather than use associational data to deduce causal relationships, why not use the description of the problem and its context?
- Idea: Use vast **knowledge base** of LLMs to extract causal chains.

LLMs and causal reasoning

SYSTEM: You are a business advisor and give clear, well-reasoned but brief (5-15 line) advice in response to questions about what people should do.

USER: A toy shop owner in the western USA wants to decide whether the ad they bought in early December is really better than their previous ads. Here is their sales data:
October: $10,200 ; November: $10,000; December: $13,000; January: $10,100
They now want to decide which ad to show in February. Can you help them decide whether the increase in sales in December was due to the ads? Note that the new ad costs $1000 more to produce, so the toy maker is interested in maximizing their profit.

GPT-4: Based on the available data, it is difficult to definitively attribute the increase in December sales to the new ad, as holiday season and other external factors may have influenced sales. To make a more informed decision, I recommend conducting an A/B test in February by dividing the targeted audience into two groups, showing one group the old ad and the other group the new ad. Compare the sales performance of both groups and calculate the net profit (considering the $1000 additional cost for the new ad) to determine which ad is more effective in driving sales and maximizing profit. This will provide a clearer understanding of the ad's impact on sales.

(a) A GPT-4 powered advisor gives sound causal advice

USER: Thanks. In February, the toy shop owner split their audience into two groups to show them two different ads. They showed the old ad to people on the newspaper; and showed the new ad to people online. When people came into the store, the owner would ask them which ad they saw. In February, they found out that they sold $6000 of toys to people who saw the new online ad, and $4500 to people who saw the old newspaper ad. What should the toymaker do in March?

GPT-4: Based on the February results, the new online ad generated $1,500 more in sales than the old newspaper ad. Considering the new ad costs $1,000 more to produce, the net profit increase is $500. In this case, it seems that the new online ad is more effective in driving sales. For March, the toy shop owner should continue using the new online ad, but also consider monitoring the long-term performance of the ad, as well as any potential changes in the target audience's preferences or external factors that may influence sales. Additionally, the owner may want to explore optimizing the new ad further to maximize its cost-effectiveness and reach.

(b) Continuing the conversation, GPT-4 gives a wrong causal answer

Using LLMs for causal discovery (Kiciman et al., 2024)

- Can we test how good LLMs are at deriving causal relationships based on a textual prompt?
- Kiciman et al. (2024) conducted a comprehensive evaluation of GPT-3.5 and GPT-4 across various causal tasks:
- Pairwise causal discovery
- Counterfactual reasoning (“what if?”)
- Token causality tasks (necessary and sufficient causes)

Example: pairwise discovery

Variable A	Variable B	Dir.
Right L1 Radiculopathy	Right adductor tendonitis	$\rightarrow$
Pharyngeal discomfort	Right C3 Radiculopahty	$\leftarrow$
Right L5 Radiculopathy	Lumbago	$\rightarrow$
Left PTA	Left L4 Radiculopahty	$\leftarrow$
Left T3 Radiculopahty	Toracal dysfunction	$\rightarrow$
DLS L5-S1	Right S1 Radiculopathy	$\rightarrow$
Left C3 Radiculopathy	DLS C2-C3	$\leftarrow$
Left C7 Radiculopathy	Left medial elbow problem	$\rightarrow$
Right Ischias	Right L5 Radiculopathy	$\leftarrow$
Right Morton trouble	Right L5 Radiculopahty	$\leftarrow$

Table 3: Example cause-effect pairs from the Neuropathic pain diagnosis benchmark. 'Dir.' refers to the ground-truth causal direction between the variables.

Figure 12: Dr. GPT

Example: counterfactual reasoning

Premise	Counterfactual Question	Multiple-choices answers
A woman does not order Chinese food.	What would have happened if she had ordered Chinese food?	The woman would have become less hungry.;The woman would have become very hungry.;That is not possible.
A woman sees a fire.	What would have happened if the woman had touched the fire?	She would have been burned.;She would not have been burned.;That is not possible.;She would have seen fire.

Figure 13: CRASS benchmark

Example: token causality

	have stopped the ball. Beyond that there was a window.				
Double preemption	Alice intends to fire a bullet at a window (AI). Bob intends to prevent Alice from hitting the window (BI). Bob tries to stop Alice (BSA). Bob is stopped by Carol (CSB). Alice fires a bullet (AF), hits the window (AH) and shatters it (WS). The window shatters (WS).	window shattering	Alice	Yes	No
Bogus preemption	Alice intents to put lethal poison into Carol's water. However, Alice does not put lethal poison into Carol's water (¬AP). Bob puts an antidote into Carol's water (BA). The water is lethal (L), if the poison is added without the addition of an antidote. If Carol would consumes the lethal water she would die (CD). Carol consumes her water (CC). Carol does not die (¬CD).	Carol's survival	Alice	No	Yes

Figure 14: Columns are (→): Vignette type, Context, Event, Actor, Necessary Cause, Sufficient Cause

Benchmark results

- Pairwise causal discovery: determining cause vs effect given two variables. GPT-4 achieved ∼97% accuracy, outperforming specialized algorithms by 13 percentage points.
    - Extension to longer chains produced moderate results; although still significantly outperformed modern algorithms.
- Counterfactual reasoning: answering "what if" questions in hypothetical scenarios. GPT-4 scored about 92% on a benchmark, ∼20 points higher than previous best.
- Causal explanation in text: analyzing vignettes to identify necessary vs sufficient causes. GPT models reached 86% accuracy in identifying causal relationships in story scenarios.
- Potential critique: memorisation. Rebuttal: authors used datasets created after the LLM's training cutoff.

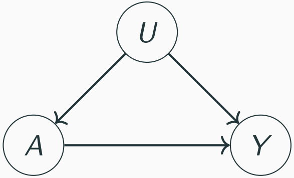
Promising?

- The above findings open possibilities for
  - an augmentation tool for policy researchers in designing causal DGPs for decision making
- However, caution is needed:
  - LLMs make unpredictable mistakes, rarely, but their semantic style makes microscopic evaluation necessary
  - Reasoning is not transparent; we must validate LLM-generated causal insights with domain knowledge or experiments.
  - Fundamental limitation: LLMs operate on correlations in text; if a causal question requires understanding beyond the data seen (e.g., physical causation), they might fail.

Confounding

Back in the world of studying people.

- Fundamental problem of causal inference is a missing data problem. $P(Y(1), Y(0))$ is never observed.
- Randomisation buys us some luck: $Y(a) \perp\!\!\!\perp A$, so that population quantities $\mathbb{E}[Y(a)] = \mathbb{E}[Y \mid A = 1]$ are identified.
- Intuition: two treatment groups are balanced (exchangeable) on average by coin-flip, so distirbution of $Y(a') \mid A$ does not change when changing a'.
    - Problem: in observation studies, we don't have control of assignment

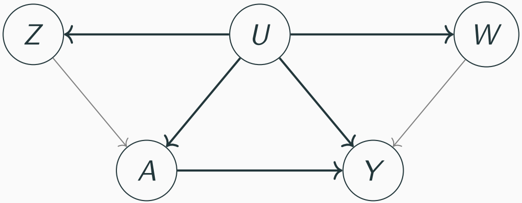


- If U is measured we can adjust, if not, we lose identification.

Proximal Causal Inference (PCI)

- **Proximal inference** (Miao et al., 2018; Eric J. Tchetgen Tchetgen et al., 2024): A framework to identify causal effects despite unobserved confounders, by using proxy variables.
- Requires two types of proxies:
  - Z: a proxy associated with the unmeasured confounder U and treatment A, but not affected by the treatment (often a pre-treatment proxy).
  - W: a proxy associated with U and outcome Y, but not affected by the outcome (often measured before Y occurs).
- Under certain assumptions (no direct $Z \to Y$ or $W \to A$ paths, and technical completeness conditions), the causal effect of A on Y can be identified via the *proximal g-formula*.
- Intuition: Z and W together contain enough information about U that we can emulate having observed U. They serve as bridges to account for confounding bias.

PCI assumptions

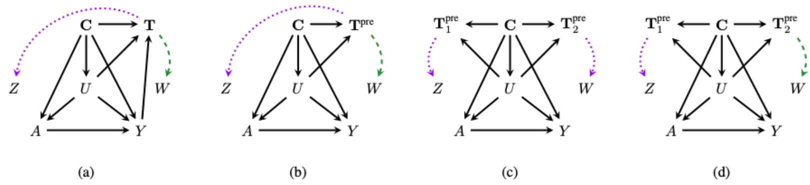


- U: unobserved confounder of treatment A and outcome Y (e.g. an underlying risk factor).
- Z: proxy for U (influenced by U, and possibly predictive of A) measured before A.
- W: proxy for U (influenced by U, and possibly by A) measured before outcome Y.
- Require: $W \perp\!\!\!\perp A \mid U, W \perp\!\!\!\perp Z \mid U, Z \perp\!\!\!\perp Y \mid A, U$
- Methods exist to recover causal effects with only one proxy but stronger assumptions are required.

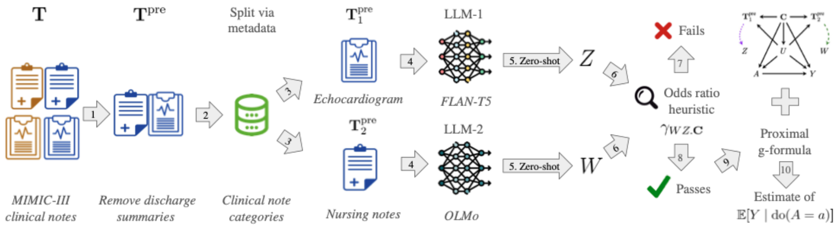
Proximal Inference with Text (Chen et al., 2024)

- Want to know if clot busting or blood thinning medication is more effective and reducing clotting.
- Measured confounders include: age, sex, blood pressure, income etc.
- Unmeasured confounder: atrial fibrillation (irregular heart rhythm) $\Rightarrow$ lose identification.
- Problem: it can be difficult to identify W and Z satisfying all required conditions
- Solution: identify Z and W which satisfy conditions by design
- Method: use rich **clinical text** (doctors' notes) that implicitly contains information about U. Idea: use LLMs as zero-shot classifiers)to infer proxies from text
  - Split each patient's pre-treatment text into two parts, T_1, T_2 (practically: ECG and nurse's notes).
  - Apply a pretrained language model on T_1 to predict the likelihood of the hidden condition U (Proxy Z).
  - Apply another (or the same) model on T_2 to independently predict U (Proxy W).

Schematic



C
T
$\mathbf{T}^{\mathrm{pre}}$
$\mathbf{T}_1^{\mathrm{pre}}$
$\mathbf{T}_2^{\mathrm{pre}}$
Z
U
W
A
Y
(a)
(b)
(c)
(d)



T
$\mathbf{T}^{\mathrm{pre}}$
Split via metadata
$\mathbf{T}_1^{\mathrm{pre}}$
Echocardiogram
$\mathbf{T}_2^{\mathrm{pre}}$
Nursing notes
LLM-1
FLAN-T5
LLM-2
OLMo
5. Zero-shot
Z
W
Fails
Odds ratio heuristic
$\gamma_{WZ.\mathbf{C}}$
Passes
Proximal g-formula
Estimate of $\mathbb{E}[Y \mid \mathrm{do}(A = a)]$
MIMIC-III clinical notes
Remove discharge summaries
Clinical note categories

Semi-synthetic experimental results

Estimation Pipeline	$(\gamma^{\text{CI low}}_{WZ.\mathbf{C}}, \gamma^{\text{CI high}}_{WZ.\mathbf{C}})$	Est. ACE	Bias	Conf. Interval (CI)	CI Cov.
P1M	$(1.35, 1.42)^{✓}$	1.304	**0.004**	(1.209, 1.394)	**Yes**
P1M, same	$(10^{16}, 10^{16})$	1.430	0.130	(1.405, 1.495)	No
P2M	$(1.82, 1.94)^{✓}$	1.343	**0.043**	(1.273, 1.425)	**Yes**
P2M, same	$(7.9, 8.41)$	1.407	0.107	(1.376, 1.479)	No

Table 1: **Fully synthetic results** with the true ACE equal to 1.3. Here, ✓ distinguishes settings that passed the odds ratio heuristic from those that failed it, with $\gamma_{\text{high}} = 2$. Corresponding to Gotcha #3, "same" indicates we used the same instance of text to infer both W and Z.

Figure 15: Results. Odds ratio heuristic is used to falsify the criterion $W \perp\!\!\!\perp Z \mid U$

Conclusion

Conclusion

- LLMs and causality can be used to study each other.
- We have seen how
  - causal inference helps us interpret model output via its architecture
  - causal philosophy improves causal reasoning
  - LLM feature extraction can remove biases from unobserved confounding
- Area is growing rapidly.

References

Chen, J. M., Bhattacharya, R., and Keith, K. A. (2024). Proximal Causal Inference With Text Data.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao (2024). An Introduction to Proximal Causal Inference. *Statistical Science*.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Kiciman, E., Ness, R. O., Sharma, A., and Tan, C. (2024). Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *Transactions on Machine Learning Research (TMLR)*.

Miao, W., Geng, Z., and Tchetgen, E. T. (2018). Identifying Causal Effects With Proxy Variables of an Unmeasured Confounder.

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, 2 edition.

Sani, N., Malinsky, D., and Shpitser, I. (2025). Explaining the Behavior of Black-Box Prediction Algorithms with Causal Learning.

Stolfo, A., Belinkov, Y., and Sachan, M. (2023). A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis.

Tang, Z., Wang, R., Chen, W., Zheng, Y., Chen, Z., Liu, Y., Wang, K., Chen, T., and Lin, L. (2025). Towards CausalGPT: A Multi-Agent Approach for Faithful Knowledge Reasoning via Promoting Causal Consistency in LLMs.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.