NeurIPS 2024:
Causality and Large Models Workshop

Teaching causal reasoning to language models

Amit Sharma

Microsoft Research

[2305.00050] Causal Reasoning and Large Language Models: Opening a New Frontier for Causality
Emre Kiciman, Robert Ness, **Amit Sharma**, Chenhao Tan.

[2407.07612] Teaching Transformers Causal Reasoning through Axiomatic Training.
Aniket Vashishtha, Abhinav Kumar, Abbavaram Gowtham Reddy, Vineeth N Balasubramanian, **Amit Sharma**

USER

Consider three variables for determining the wetness of grass.
Rain (R)
Sprinkler is on (SO)
Wet Grass (WG)

Construct a causal graph between these variables.

Wait, this example is from Pearl's causality book. Likely that it was ***in GPT-4 training set and has been memorized.***

GPT-4

Given the usual relationships between these variables:
1. **Rain can cause the grass to be wet.**
2. If the sprinkler is on, it can cause the grass to be wet.

Based on this:
1. Both Rain (R) and Sprinkler is on (SO) can be considered as parent nodes, because they can influence the status of the Wet Grass (WG).
2. Wet Grass (WG) can be considered as a child node, since its status can be influenced by either Rain (R) or Sprinkler is on (SO).

Here is a simple representation of the causal graph:

R --> WG
SO --> WG

Where:
* R points towards WG indicating that rain can cause the grass to be wet.
* SO points towards WG indicating that if the sprinkler is on, it can make the grass wet.
Note: It's important to mention that in a real-world scenario, there can be other confounding variables or interactions that could influence this simple relationship, but based on the variables provided, the above graph is appropriate.

USER

Consider three variables for determining the wetness of grass **grown indoors in a closed room.**
Rain (R)
Indoor Sprinkler is on (SO)
Wet Indoor Grass (WG)

Construct a causal graph between these variables.

GPT-4

Based on the information given:
1. **Rain (R) cannot directly cause the grass grown indoors in a closed room to be wet** because the rain is outside and cannot penetrate the room. So there's no direct arrow from R to WG.
2. If the Indoor Sprinkler (SO) is on, it can cause the Indoor Grass (WG) to be wet. So there is a direct arrow from SO to WG.
3. **Rain might influence whether someone decides to turn on the indoor sprinkler**, assuming that one might not water the indoor grass if it's raining outside, thinking that the moisture or humidity might increase indoors. So, there could be a direct arrow from R to SO.
The causal graph would then look like:

R → SO → WG

This graph indicates:
* Rain influences whether the indoor sprinkler is turned on.
* The indoor sprinkler being on causes the indoor grass to be wet.
However, keep in mind that this is a simplified model. In reality, there may be other factors at play, and the relationship between rain and the decision to turn on an indoor sprinkler is an assumption made for this example.

GPT-4 is not reliable, but it is not bad either.
How generalizable is this capability to infer causal graphs?

Across scientific fields, LLMs are great at determining causal relationships

Pairwise causal direction:

- 96% accuracy on Neuropathic Pain and Tubingen Pairs datasets

Full graph discovery

- Higher accuracy than discovery algorithms that use data
  - Arctic Sea Ice, Covid-19, Alzheimers
  - Also valid for graphs released after LLM's training cutoff date

You are a helpful assistant to a neuropathic pain diagnosis expert. Which cause-and-effect relationship is more likely?

A. Left T6 Radiculopathy causes DLS T5-T6.
B. DLS T5-T6 causes Left T6 Radiculopathy.
C. No causal relationship exists.

Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.

[TMLR] Causal Reasoning and Large Language Models: Opening a New Frontier for Causality
Emre Kiciman, Robert Ness, **Amit Sharma**, Chenhao Tan.

Different kinds of causal reasoning

Premise: *Suppose there is a closed system of 4 variables, A, B, C and D. All the statistical relations among these 4 variables are as follows: A correlates with B. A correlates with C. A correlates with D. B correlates with C. B correlates with D. C correlates with D. However, B and D are independent given A. B and D are independent given A and C. C and D are independent given A. C and D are independent given A and B.*

Hypothesis: There exists at least one collider (i.e., common effect) of A and B.

Label: *No.*

Infer Causal Graph from Correlational Statements: Corr2Cause [Jin et al. 23]

Question: Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships:

Physical vulnerability has a direct effect on the likelihood of **fatality** and **vaccination decision**. **Vaccination** has a direct effect on the **fatality rate**.

In the entire population, 50% of the people are vulnerable to a certain disease.
For vulnerable and vaccinated people, the fatality rate is 4%. For vulnerable and unvaccinated people, the fatality rate is 7%.
For strong and vaccinated people, the fatality rate is 1%. For strong and unvaccinated people, the fatality rate is 5.8%.
Overall, the fatality rate for vaccinated people is 5%, while the fatality rate for unvaccinated people is 4.5%.

Does getting vaccinated increase the likelihood of death?

Ground-Truth Answer: No

Estimate causal quantities given the graph: CLadder [Jin et al. 23]

Premise: *Suppose there is a closed system of 4 variables, A, B, C and D. All the statistical relations among these 4 variables are as follows: A correlates with B. A correlates with C. A correlates with D. B correlates with C. B correlates with D. C correlates with D. However, B and D are independent given A. B and D are independent given A and C. C and D are independent given A. C and D are independent given A and B.*

Hypothesis: There exists at least one collider (i.e., common effect) of A and B.

Label: *No.*

Question: Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships:

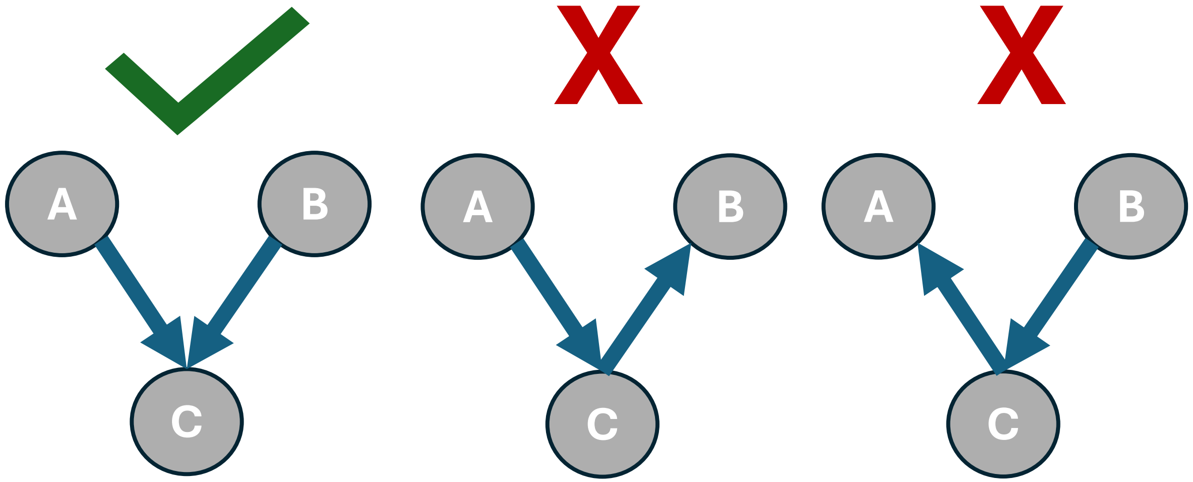
Physical vulnerability has a direct effect on the likelihood of **fatality** and **vaccination decision**. **Vaccination** has a direct effect on the **fatality rate**.

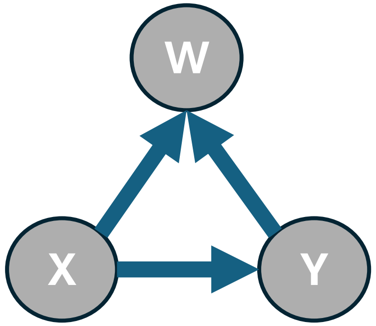
In the entire population, 50% of the people are vulnerable to a certain disease.
For vulnerable and vaccinated people, the fatality rate is 4%. For vulnerable and unvaccinated people, the fatality rate is 7%.
For strong and vaccinated people, the fatality rate is 1%. For strong and unvaccinated people, the fatality rate is 5.8%.
Overall, the fatality rate for vaccinated people is 5%, while the fatality rate for unvaccinated people is 4.5%.

Does getting vaccinated increase the likelihood of death?

Ground-Truth Answer: No

CLadder





Requires understanding of a Collider

(e.g., graph reachability, d-separation)

Requires understanding of effect identification

(e.g., d-separation, do-calculus)

LLMs such as GPT-4 do not do well on these tasks.

Can we teach causal reasoning to a language model?

USE AS A **VERIFIER** OF LLM OUTPUT (CAUSAL REWARD MODEL)

DEVELOP **INDUCTIVE BIASES** TO IMPROVE REASONING WHILE TRAINING LANGUAGE MODELS

Key insight: Causal reasoning can be broken down into fundamental axioms

High-level tasks such as graph discovery, effect inference, attribution

depend on

A small number of axioms such as transitivity, d-separation, backdoor criterion, do-calculus rules, etc.

Rather than training on individual tasks, **teach a language model axioms of causality [Axiomatic Training].**

If a model learns the axioms and how to compose them, it can solve any causal task!

True for other formal reasoning systems too (e.g., logic, math)

We will consider two axioms for model to learn: transitivity and d-separation

Transitivity (graph reachability)

If A causes B, and B causes C, then **A can cause C; C cannot cause A.**

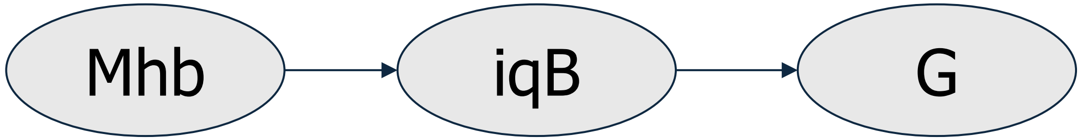
d-separation (causal independence)

X and Y are *d-separated* by Z if all paths between any node in X and any node in Y are blocked by the conditioning set Z.

A path between X and Y is *blocked* by Z if there exists a node $A \in Z$

- A is the parent node in a fork structure on the path (i.e., $\cdot \leftarrow A \rightarrow \cdot$);
- A is the mediator node in a chain structure on the path (i.e., $\cdot \rightarrow A \rightarrow \cdot$);
- In any collider structure on the path (i.e., $\cdot \rightarrow A \leftarrow \cdot$), Z does not contain A or its descendants.

Axiom is learnt if language model can apply the axiom multiple times to obtain answer



Mhb causes iqB. iqB causes G. Does G cause iqB?: No

X causes Tubc. Tubc causes Bb. Tubc causes WYTi. WYTi causes R7. R7 causes M. Bb causes R7. R7 causes CiFQ. Does Tubc cause CiFQ?: Yes

Since the graph structure is arbitrary, cannot be solved unless a model *applies* the axiom multiple times.

Eval	w3 causes ROv. w3 causes tQC. H causes ROv. H causes tQC. b causes ROv. b causes w3. b causes H. Does tQC cause ROv?: No	Branching
Eval	LKk causes 5Ov. Kk causes L0. L0 causes KWO. 5Ov causes c. Does KWO cause L0?: No	Shuffled Sequences
Eval	FDAH26mV7 causes 7tzaIHjlY. 7tzaIHjlY causes 0kspcX95Im. 0kspcX95Im causes 7rhFSlx2o9. 7rhFSlx2o9 causes 1PlG5LHVqp. Does FDAH26mV7 cause 7tzaIHjlY?: Yes	Sequences with Longer Node Names
Eval	r causes rZ. rZ causes L. L causes bUx. bUx causes Pbr. Pbr causes 1w. 1w causes c3. c3 causes yBQ. yBQ causes yK. yK causes w. w causes P. P causes kH. kH causes 1u. 1u causes jV7. jV7 causes i. Does r cause rZ?: Yes	Long Linear Sequences
Eval	rU6 causes eF. eF causes ivC. 3R causes ivC. 3R causes A8. 2 causes A8. 2 causes i. i causes a03. y causes a03. b causes y. b causes h. h causes yN. ic0 causes yN. ic0 causes Hd. Hd causes U. Does rU6 cause eF?: Yes	Long Sequences with Random Flipping

Let's evaluate existing LLMs and see what happens!

Since the graph structure is arbitrary, cannot be solved unless a model *applies* the axiom multiple times.

Evaluating language models

Models

- GPT-4 *(1.5 T params)*
- Gemini Pro *(~100 B params)*
- Phi-3 *(3.8 B params)*

Two settings

- Zero-shot
- With in-context examples

Metric: Accuracy

Evaluating Transitivity / Graph Reachability

X causes Tubc. Tubc causes Bb. Tubc causes WYTi. WYTi causes R7. R7 causes M. Bb causes R7. R7 causes CiFQ. Does Tubc cause CiFQ?: Yes

Corr2Cause Dataset

Premise: Suppose there is a closed system of 5 variables, A, B, C, D and E. All the statistical relations among these 5 variables are as follows: A correlates with C. A correlates with D. A correlates with E. B correlates with C. B correlates with D. C correlates with D. C correlates with E. D correlates with E. However, A is independent of B. A and B are independent given E. A and D are independent given B and C. A and D are independent given B, C and E. A and D are independent given C. A and D are independent given C and E. B and D are independent given A and C. B and D are independent given A, C and E. B and D are independent given C. B and D are independent given C and E. B is independent of E. B and E are independent given A. B and E are independent given A and C. B and E are independent given A, C and D. B and E are independent given A and D. C and E are independent given A. C and E are independent given A and B. C and E are independent given A, B and D. C and E are independent given A and D. D and E are independent given A. D and E are independent given A and B. D and E are independent given A, B and C. D and E are independent given A and C. D and E are independent given B and C. D and E are independent given C.

Hypothesis: E is a cause for D, but not a direct one.

Large pre-trained language models fail on these tasks, except GPT-4 on transitivity

Model	7	10	15
Phi-3	0.85	0.89	0.85
Gemini Pro	0.73	0.81	0.66
GPT-4	0.98	0.87	0.86
w/ in-context			
Phi-3	0.92	0.87	0.82
Gemini Pro	0.82	0.79	0.78
GPT-4	**0.99**	**0.94**	**0.94**

Transitivity Test Set: Causal chains of different lengths

Model	Acc
Phi-3	0.52
Gemini Pro	0.52
GPT-4	0.58
w/ in-context	
Phi-3	0.58
Gemini Pro	0.52
GPT-4	**0.64**

Corr2Cause Test Set: Questions with number of nodes = 5

Let’s try training a tiny model (67M) with **axiomatic training**

Axiomatic Training: Learn from demonstrations of an axiom

Training Example structure

Premise. Hypothesis. Label.

Mhb causes iqB. iqB causes G. Does G cause iqB?: No

M causes A. A causes CG. Does M cause CG?: Yes

B causes A. C causes As. A causes C. Does B cause As?: Yes

X causes Tubc. Tubc causes Bb. Tubc causes WYTi. WYTi causes R7. R7 causes M. Bb causes R7. R7 causes CiFQ. Does Tubc cause CiFQ?: Yes

Train set

3-6 length sequences

Test set (same as before)

7-15 length sequences

Key to Generalization:
Sufficient diversity in training set

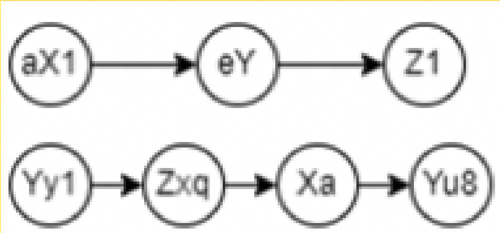
Variability in training set

1. Length Level: Chains of length 3 to 6 nodes
2. Node Level: Name length range 1-3 characters
3. Edge Level:
   a. Sequential : $X \rightarrow Y \rightarrow Z$
   b. Random Flipping: $X \rightarrow Y \leftarrow Z$

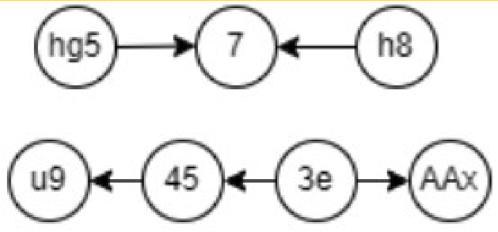
Summary: Training and Evaluation Setup

TRAIN SET

Sequential Chains



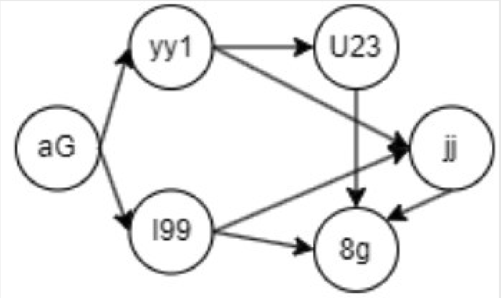
Chains with Random Flipping



Transformer

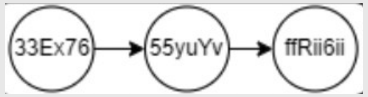
EVAL SET

Branching

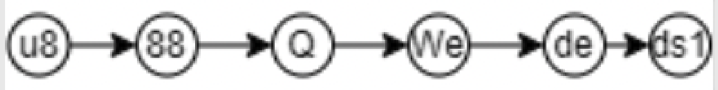


Reversal

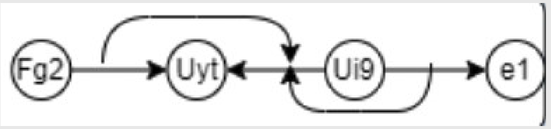
Longer Node Names



Longer Causal Chains



Shuffling

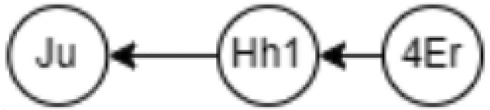


Model architecture and Training

- **67M** Decoder-only model based on GPT-2
- Trained **from scratch**
- Custom **tokenizer**
  - Special tokenizer for “causes”
  - Character-level tokenizer for variable names
- **Positional Encoding:** None (see paper for ablations)
- Train dataset size: **175k examples**
- Trained for **100 epochs**

Results: Similar to GPT-4 on sequence chains, much more accurate than billion-scale models such as Gemini Pro and Phi-3

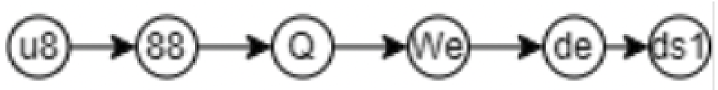
Reversal



Model	3	4	5	6
Baselines				
Zero Shot				
GPT-4	0.97	<u>0.99</u>	0.98	0.92
Gemini Pro	0.61	0.59	0.66	0.62
Phi-3	0.80	0.69	0.73	0.69
Multi Shot				
GPT-4	**1.00**	**1.00**	**1.00**	**0.99**
Gemini Pro	0.95	0.87	0.77	0.71
Phi-3	0.93	0.89	0.75	0.75
Axiomatic Training				
TS1 w NoPE	0.98	0.99	0.92	0.91

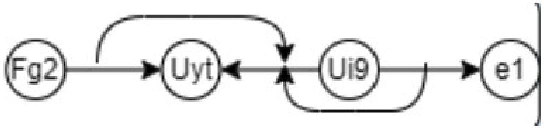
Results: Similar to GPT-4 on sequence chains, much more accurate than billion-scale models such as Gemini Pro and Phi-3

Longer Causal Chains



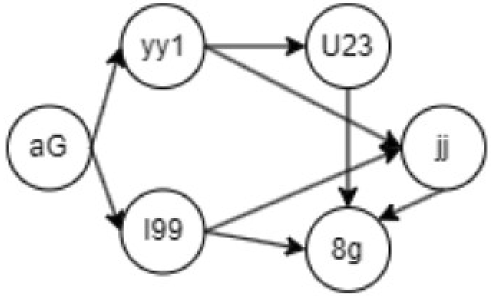
Length	7	8	9	10	11	12	13	14
GPT-4	**0.98**	**0.93**	**0.94**	**0.87**	**0.95**	**0.92**	**0.93**	**0.93**
Ours	0.98	0.97	0.91	0.90	0.92	0.90	0.83	0.84

Shuffling



Length	3	4	5	6	7	8
GPT-4	**0.99**	**0.97**	**0.89**	**0.85**	**0.95**	**0.90**
Ours	1.00	0.95	0.87	0.84	0.79	0.76

Branching



Length	5 (BF =2)	8 (BF=2)	10 (BF=2)	12 (BF=2)
GPT-4	**0.99**	**0.89**	**0.85**	**0.95**
Ours	0.83	0.86	0.69	0.64

Results: Corr2Cause dataset

- Same 67M model architecture.
- **Trained** on a subset of Corr2Cause dataset of graph size= {3,4}
- **Test:** Graph size 5

Model	Precision	Recall	F1 Score	Accuracy
Ours	**0.72**	0.50	0.59	**0.64**
Zero-Shot				
Phi-3	0.52	0.60	0.56	0.52
Gemini pro	0.52	0.59	0.55	0.52
GPT-4	0.59	0.50	0.54	0.58
Multi-Shot				
Phi-3	0.57	0.67	0.61	0.58
Gemini pro	0.51	**0.74**	0.60	0.52
GPT-4	0.66	0.56	**0.61**	**0.64**

Extensions and Future Work

- **Backdoor criterion:** Extending it to learn backdoor criterion
- **Multiple axioms:** Learn multiple axioms and test whether LM can compose them
- **Beyond causality:** Can axiomatic training work with propositional logic?
- **Natural data:** Extending to natural language inputs

"Adam dropped the glass vase. The vase hit the floor. It broke." | ***A->B ->C***
"Adam dropped the glass vase. It broke." | ***A->C***

The way forward:
Expanding the applicability of causal methods

Causal Questions (with data)

- Use LLMs to parse text to formal causal question
  - Human intent
  - Scientific domain knowledge
- Use causal methods to answer the question

(Informal) Causal Questions:

  - Use LLMs to answer questions
  - Use causal axioms to verify