

Causality & large language models: A new frontier

Amit Sharma

Microsoft Research India

@amt_shrma

www.amitsharma.in

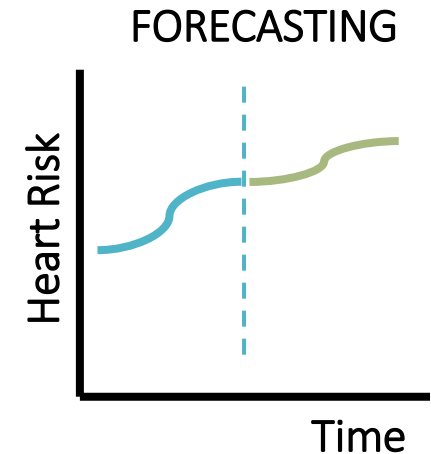
Joint work with **Amit Deshpande**, Emre Kiciman, Abhinav Kumar, Robert Ness & Chenhao Tan

Causality: Study of action and effect



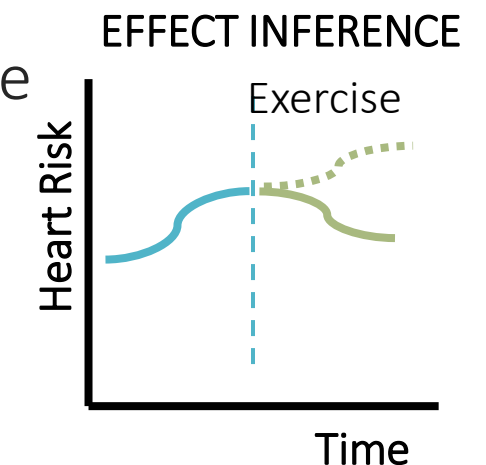
(Correlational) Prediction: Based on past health data, what is the chance of a heart attack in the next 2 years?

- **Forecasting** patterns in available data



(Causal) Prediction: What should I do today to reduce the chance of a heart attack?

- **Discovering** the factors that impact heart health (e.g., exercise)
- No past data on exercise for this person
- Estimate based on **causal effect** of exercise

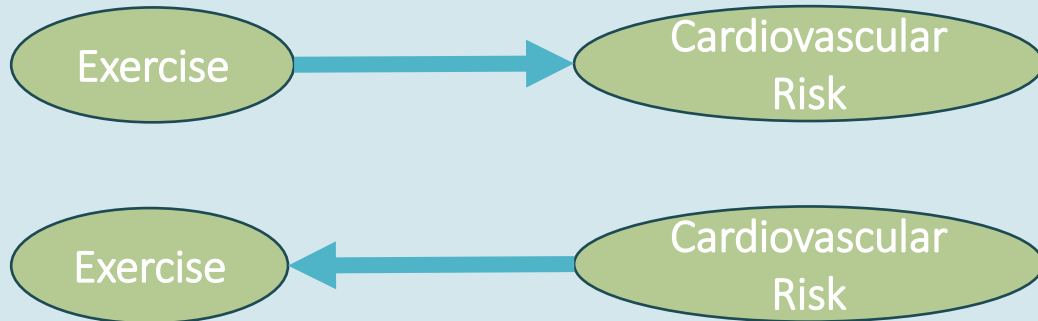


Two challenges in causal inference

“What should I do today to reduce the chance of a heart attack?”

1

Infer causal relationships



2

Reason over unstructured data

A lovely xx-year-old gentleman with multiple issues for comprehensive evaluation.

1. *Tinnitus* of many years' duration. He has no *neurologic complaints*, no headache, no *vertigo*, no sudden changes in hearing.

2. *Low-grade pancreatic neoplasm*.

In xxxx, presented with *epigastric pain* and high CA 19-9. CT showed a mass in the *pancreas*.

So far, answering these questions has been a difficult problem.

Most studies assume that this information is provided by a human domain expert.

As a result, causal analysis is considered “impractical” for many real-world scenarios.

Can large language models (LLMs) address these challenges?

- Infer causal relationships
- Reason over unstructured data

Today's talk: Can large language models make causal inference practical?

Two fundamental challenges for causal inference (require human input)

- Learning causal relationships
- Reasoning over unstructured data (e.g., text)

LLMs are good in exactly these tasks that need human input.

- Results on causal discovery
- Results on counterfactual inference from text
- But LLMs have unpredictable errors

Future: How to combine LLM + principled causal methods?

- Big opportunity to accelerate causal analysis across scientific fields

Today's results on LLMs are equally applicable to a wide variety of domains



What is the effect of a treatment on patients' outcomes?



Why did this customer discontinue their product subscription?



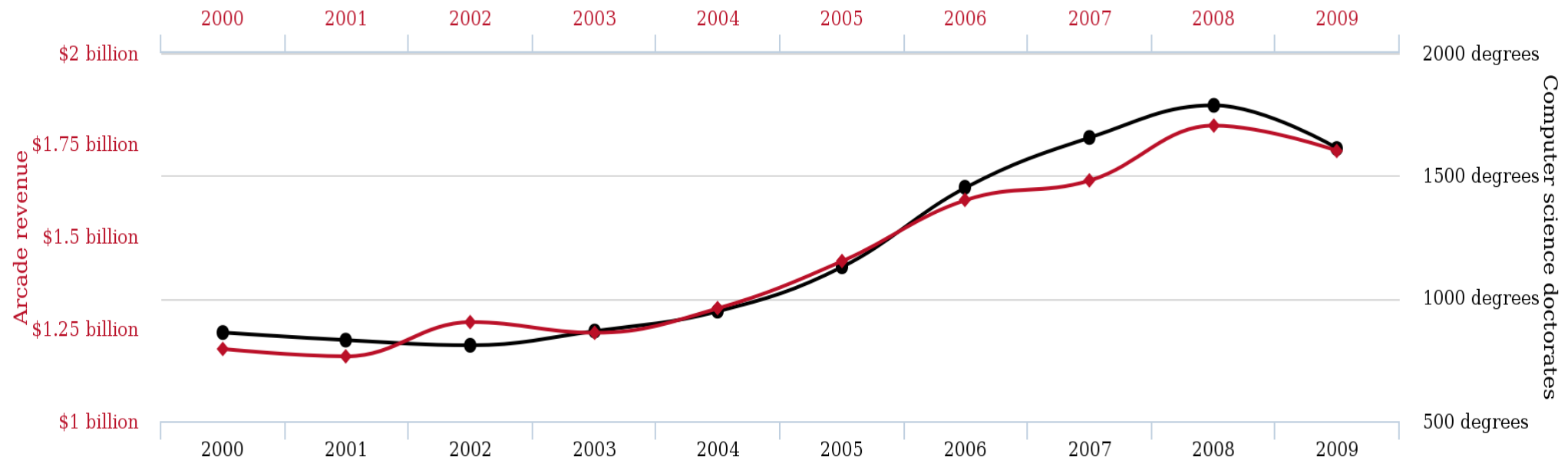
Do people influence their friends' preferences in music? In politics?



Would this government regulation lead to a decrease in air pollution?

I. LLMs capabilities in graph discovery (Challenge 1)

Fundamental challenge 1: Learning causal relationships from observed data is hard



tylervigen.com

<http://www.tylervigen.com/spurious-correlations>

Provably impossible to recover causal direction between two variables

Observed distribution: $P(A,B)$

- Graph 1 factorization: $P(B|A) P(A)$

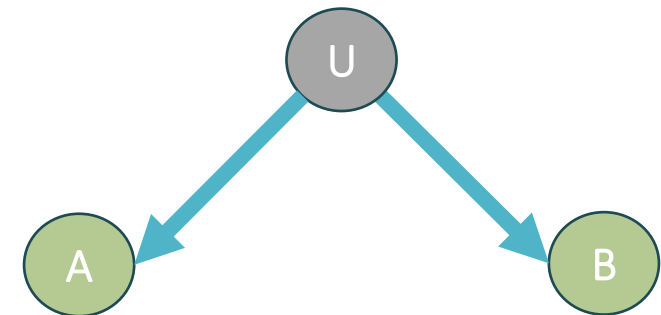
- Graph 2 factorization: $P(A|B) P(B)$

Same likelihood under both graph models.

Impossible to determine direction non-parametrically.

In practice, there can be **unobserved variables**.

- Even harder to detect such confounding.



Full graph discovery admits some testable graph structures,
but the equivalence class of possible graphs is still large enough.

In last two decades, many algorithms for graph learning in specific settings

Faster tests for special graph structures that lead to different likelihood (KCIT, RCIT) [Strobl et al. 2017]

Linear models with non-gaussian noise, or non-linear models (LinGAM) [Shimizu et al. 2006]

Formulate as a continuous optimization problem and use deep learning (NOTEARS) [Zhang et al. 2018]

But still not enough!

Sobering results using SoTA graph discovery algorithms on real-world data from domains like medicine and atmospheric science due to the difficulty of learning causal relationships from data alone.

[Tu et al. 2019, Huang et al. 2021, Kaiser & Sipos 2022]

Large language models (LLMs)

Transformer-based generative language models such as ChatGPT, GPT-4

Trained on next token prediction on large amounts of text

+ Finetuning on following instructions [Optional]

We study models from **2.7B** parameters (Ada) to **175B+** (GPT-3.5, GPT-4)

As model scale increases, many “emergent” abilities in language understanding, programming, creative writing, etc.

What about causality?



USER

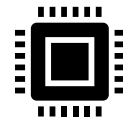
Consider three variables for determining the wetness of grass.

Rain (R)

Sprinkler is on (SO)

Wet Grass (WG)

Construct a causal graph between these variables.



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:

$R \rightarrow WG$

$SO \rightarrow WG$

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.

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



USER

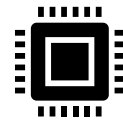
Consider three variables for determining the wetness of grass **grown indoors in a closed room.**

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:

$R \rightarrow SO \rightarrow WG$

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.

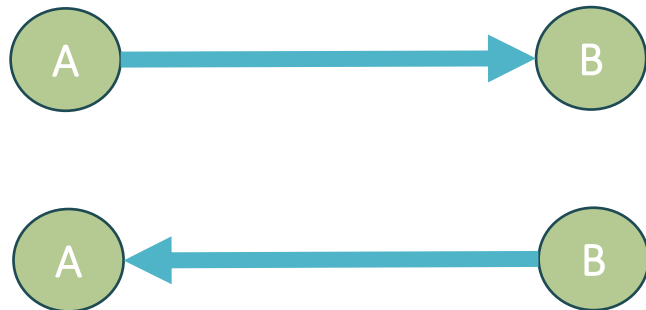
GPT-4 is not perfect, but it is not bad either.

*How robust is this capability to infer causal graphs?
And how is it doing this?*

Let's try it out on a wide variety of domains

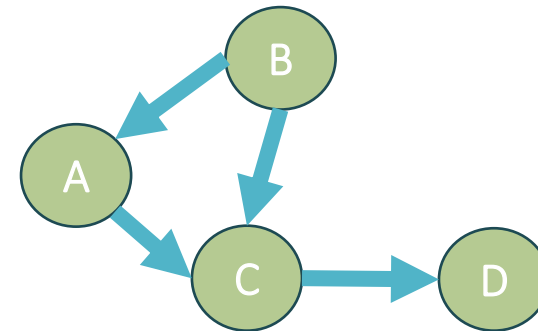
PAIRWISE DISCOVERY

Given a pair of variables (A,B),
decide whether A causes B or B
causes A?



FULL GRAPH DISCOVERY

Given a set of variables, infer a
directed acyclic graph over them.
Infer which pairs of variables form an
edge, and their direction.



Pairwise discovery: Tübingen Benchmark

104 variable pairs spanning range of fields [Mooij et al. 2016]

Variable A	Variable B	Domain
Age of Abalone	Shell weight	Zoology
Cement	Compressive strength of concrete	Engineering
Alcohol	Mean corpuscular volume	Biology
Organic carbon in soil	Clay content in soil	Pedology
PPFD (Photosynthetic Photon Flux Density)	Net Ecosystem productivity	Physics
Drinking water access	Infant mortality	Epidemiology
Ozone concentration	Radiation	Atmospheric Science
Contrast of tilted Gabor patches	Accuracy of detection by participants	Cognitive Science
Time for 1/6 rotation of a Stirling engine	Heat bath temperature	Engineering
Time for passing first segment of a ball track	Time for passing second segment	Basic Physics

Challenging task: Most existing algorithms achieve 70-80% accuracy, Best is 83% [Mosaic, Wu & Fukumizu 2020].

Applying LLMs to pairwise causal discovery

Method: For each pair, input below prompts to an LLM and record the output.

TWO PROMPTS PER PAIR

Template:

- Does changing {A} cause a change in {B}? Please answer in a single word: yes or no.
- Does changing {B} cause a change in {A}? Please answer in a single word: yes or no.

Examples:

- Does changing the altitude cause a change in temperature? Please answer in a single word: yes or no.
- Does changing the temperature cause a change in altitude? Please answer in a single word: yes or no.

SINGLE PROMPT

Template:

- Which cause-and-effect relationship is more likely?
A. changing {A} causes a change in {B}.
B. changing {B} causes a change in {A}.

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>`.

Example:

- Which cause-and-effect relationship is more likely?
A. changing the altitude causes a change in temperature.
B. changing the temperature causes a change in altitude.

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>`.

SYSTEM: You are a helpful assistant for causal reasoning.

USER: Which cause-and-effect relationship is more likely?

A. changing the age of an abalone causes a change in its length.

B. changing the length of an abalone causes a change in its age.

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>.

GPT-3.5-TURBO: 1. Cause: Changing the age of an abalone

Effect: a change in its length.

2. Cause: Changing the length of an abalone

Effect: a change in its age.

Based on the biological knowledge about abalones, it is more likely that changing the age of an abalone causes a change in its length. This is because the length of an abalone is known to increase as it ages, and not the other way around. Therefore, the correct answer is <Answer>A</Answer>.

(a) Correct answer

Figure 3: **Probing reasoning of LLMs for deciding causal direction.** Two example outputs from a LLM (gpt-3.5-turbo). The first gives the impression of the LLM performing causal reasoning to arrive at the correct answer—that age may cause the length of abalone to increase. The second example asks a question that requires arguably the same causal knowledge but the LLM proceeds with an incoherent argument and provides an incorrect answer.

Results: GPT-4 obtains 96% accuracy, 13 points higher than previous SoTA

Evaluate all available models from OpenAI.

Two prompts template:

- With increased model size (and perhaps RLHF), begin to see gains.
 - Smaller models (ada, babbage, curie) are similar to random chance.
- With a single system prompt change, obtain **5 points gain!**
 - System prompt: “You are a helpful assistant for causal reasoning”

Single prompt template:

- Obtain further gains in accuracy
- GPT-4 obtains 97%
 - Surprising how few LLM errors are!

Model	Acc.	Wt. Acc.
Slope (Marx & Vreeken, 2017)	0.75	0.83
bQCD (Tagasovska et al., 2020)	0.68	0.75
PNL-MLP (Zhang & Hyvarinen, 2012)	0.75	0.73
Mosaic (Wu & Fukumizu, 2020)	83.3	81.5

Covariance-based

Knowledge-based

Having clear, specific prompts matters

Input Pair: Ozone concentration, Radiation

- LLM: 1. Ozone is a gas in Earth's atmosphere that absorbs most of the Sun's ultraviolet (UV) radiation.
2. Chlorofluorocarbons (CFCs), pollutants that were once widely used in refrigerants and aerosol sprays, have been found to destroy ozone molecules in the atmosphere.
3. The destruction of the ozone layer in the upper atmosphere causes more UV radiation to reach the surface of Earth.
4. Increased UV radiation can lead to various negative effects such as skin cancer, cataracts, and damage to crops.

From these facts, we can determine that a decrease in ozone concentration would lead to an increase in radiation. Therefore, Option A is more likely:

<Answer>A</Answer>

USER: Wait, this question is not about stratospheric ozone concentration. Rather, I'm asking about the ground-level ozone concentration measured in a city and the corresponding ground-level radiation due to sunlight.

GPT-3.5-TURBO:: I see, thank you for clarifying. In that case, the answer is likely option B: changing radiation causes a change in ground-level ozone concentration. This is because ground-level ozone is formed through a chemical reaction between nitrogen oxides (NOx) and volatile organic compounds (VOCs) in the presence of sunlight. This reaction is known as photochemical smog and occurs more frequently on hot, sunny days when there is more radiation. Therefore, an increase in radiation would likely lead to an increase in ground-level ozone concentration.

<Answer>B</Answer>

Similar results on Neuropathic Pain dataset

221 nodes & 475 edges about neuropathic pain diagnosis [Tu et al. 2019]. Use all edges as pairs.

Variable A	Variable B	Dir.	Model	Accuracy
Right L1 Radiculopathy	Right adductor tendonitis	→	ada	40.1
Pharyngeal discomfort	Right C3 Radiculopahty	←	text-ada-001	50.0
Right L5 Radiculopathy	Lumbago	→	babbage	50.0
Left PTA	Left L4 Radiculopahty	←	text-babbage-001	50.9
Left T3 Radiculopahty	Toracal dysfunction	→	curie	50.0
DLS L5-S1	Right S1 Radiculopathy	→	text-curie-001	50.0
Left C3 Radiculopathy	DLS C2-C3	←	davinci	38.4
Left C7 Radiculopathy	Left medial elbow problem	→	text-davinci-001	50.0
Right Ischias	Right L5 Radiculopathy	←	text-davinci-002	51.7
Right Morton trouble	Right L5 Radiculopathy	←	text-davinci-003	55.1
			gpt-3.5-turbo	71.1
			gpt-3.5-turbo (neuropathic pain expert)	75.1
			gp4-4	78.4
			gpt-4 (neuropathic pain expert)	84.3
			text-davinci-003 (single prompt)	86.0
			gpt-3.5-turbo (single prompt)	85.5
			gpt-4 (single prompt)	96.2

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

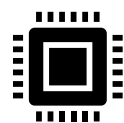


USER

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.

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>.



GPT-4

First, we need to understand the terms involved in the cause-and-effect relationship.

Left T6 Radiculopathy refers to a condition where there is damage or irritation to the nerve root that exits the spinal cord at the level of T6 on the left side of the body.

DLS T5-T6 refers to degenerative changes in the discs, ligaments, and/or bones at the level of T5-T6 in the spine.

Now, let's consider the two options:

A. Left T6 Radiculopathy causes DLS T5-T6. This would mean that the damage or irritation to the nerve root at T6 on the left side of the body somehow leads to degenerative changes in the discs, ligaments, and/or bones at the level of T5-T6. While it is possible that nerve damage could contribute to degenerative changes in the spine, it is more likely that the opposite is true.

B. DLS T5-T6 causes Left T6 Radiculopathy. This option suggests that the degenerative changes in the discs, ligaments, and/or bones at the level of T5-T6 are causing the Left T6 Radiculopathy. This is a more likely cause-and-effect relationship, as degenerative changes in the spine can lead to nerve root irritation or compression.

Therefore, the more likely cause-and-effect relationship is: <Answer>B</Answer>DLS T5-T6 causes Left T6 Radiculopathy.

Example: LLM correctly describes medical terms and can identify the causal direction.

Example: “Incorrect”, but plausible LLM answer

Input Pair: L5 Radiculopathy, Obesity

Neuropathic Pain dataset: L5 Radiculopathy causes Obesity.

LLM (gpt-3.5-turbo): Obesity causes L5 Radiculopathy.

Medical Literature: Obesity may be a risk factor for radiculopathy [Atchison & Vincent, 2012; Tomic et al., 2009]

LLM: Step 1: Understand the terms

- Neuropathic pain: pain caused by damage or disease affecting the nervous system
- L5 Radiculopathy: a condition where the nerve root that exits the spine at the L5 level is compressed or irritated, causing pain, numbness, or weakness in the leg
- Obesity: a medical condition characterized by excessive body fat that increases the risk of health problems

Step 2: Analyze the cause-and-effect relationship

- Option A: Left L5 Radiculopathy causes Left Obesity. This relationship does not make sense because L5 radiculopathy is a nerve condition that affects the leg, while obesity is a whole-body condition related to excessive body fat. It is unlikely that a leg nerve condition would cause obesity.

- Option B: Left Obesity causes Left L5 Radiculopathy. This relationship is more plausible because obesity can put extra pressure on the spine and nerves, leading to nerve compression or irritation. It is possible that obesity could cause L5 radiculopathy.

Step 3: Choose the more likely cause-and-effect relationship

- Based on the analysis, Option B is more likely: Left Obesity causes Left L5 Radiculopathy.

<Answer>B</Answer>

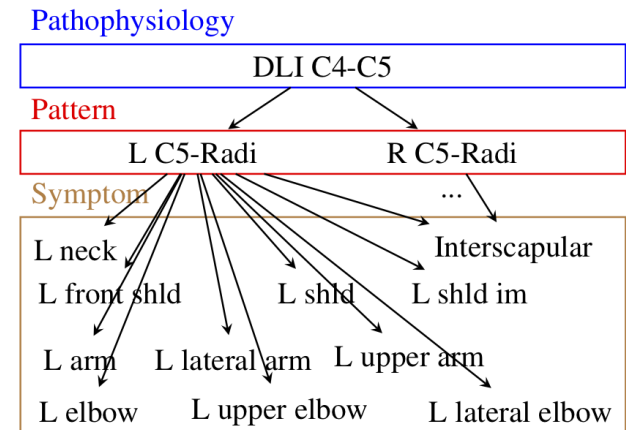
Full graph discovery: Neuropathic pain dataset

Use a 100 pair subsample [Tu et al. 2023]

50 correct edges, 50 non-existent edges

Need to infer existence + direction of an edge

Ground-truth graph



Single prompt

Template:

- Which cause-and-effect relationship is more likely? Consider only direct causal mechanism and ignore any effect due to common causes.

- A. {A} causes {B}.
- B. {B} causes {A}.
- 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>.

Example:

- Which cause-and-effect relationship is more likely? Consider only direct causal mechanism and ignore any effect due to common causes.

- A. Left L4 Radiculopathy causes Right Iliac Crest pain.
- B. Right Iliac Crest pain causes Left L4 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>.

Prompt matters! With an improved prompt, we obtain 3x better accuracy in graph discovery

Model	Precision	Recall	F1
Random	0.25	0.5	0.33
chatGPT (Tu et al., 2023)	1	0.12	0.21

Simple use of ChatGPT resulted in **0.21 F1** on retrieving edges correctly.

With our single prompt, gpt-3.5-turbo (ChatGPT) obtains **0.68 F1**.

On an atmospheric science dataset, LLMs obtain higher accuracy than recent deep learning algorithms

Arctic sea ice dataset: 12 variable, 48 edges graph on the drivers of sea ice thickness in the Arctic region.

Variables: Total cloud water path, sea level pressure, geopotential height, meridional and zonal wind at 10m, net shortwave and longwave flux at surface, etc.

Algorithm	NHD	No. of predicted edges	Baseline NHD	Ratio
TCDF	0.33	9	0.39	0.84
NOTEARS (Static)	0.33	15	0.44	0.75
NOTEARS (Temporal)	0.35	7	0.38	0.92
DAG-GNN (Static)	0.32	23	0.49	0.65
DAG-GNN (Temporal)	0.34	16	0.44	0.77
gpt-3.5-turbo	0.33	62	0.76	0.43
gpt-4	0.22	46	0.65	0.34

Table 7: Normalized hamming distance (NHD) for different causal discovery algorithms. Since NHD depends on the number of predicted edges, we compare the ratio of NHD and baseline NHD across algorithms. A lower NHD ratio is better. LLM-based discovery (gpt-3.5-turbo) obtains comparable NHD and the lowest NHD ratio compared to recent covariance-based discovery algorithms.

Construct Validity: Has GPT-4 memorized the benchmarks?



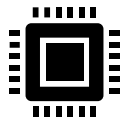
SYSTEM

You are an AI assistant that has read many sources of text from the internet. I am looking at text from the dataset, published by as . Here is the README for the dataset: I am going to list some sample rows of data, and I want you to complete each row as best as possible. I am testing your memory.



USER

pair0005 Age



GPT

pair0005 Age Length Abalone ->

	Cells	Rows
GPT-3.5	58.9%	19.8%
GPT-4	61%	25%

Yes, Tübingen dataset clearly in the training dataset.

Construct Validity: What are we measuring?

Let us model knowledge-based discovery as:

$$\underbrace{P(Y|D)}_{\text{Likelihood that knowledge can be transformed to answer question } Y} \underbrace{P(D)}_{\text{Likelihood that knowledge } D \text{ is known by LLM}}$$

Conservative interpretation:

With memorized benchmark data, we are *not* measuring $P(D)$.

We *are* measuring: how LLM can process and transform D into the necessary causal relationship Y

Another interpretation:

Even the most popular dataset, Tübingen, is not fully memorized.

There is still some non-trivial understanding of concepts and reasoning over them.

Implications for causality in practice:

A useful tool to accelerate causal studies

All downstream causal analysis depend on the causal graph.

Given a problem, LLMs can

-- **Generate a candidate graph**, which can be validated by an expert.

OR

-- **Critique a causal graph** developed by experts or discovery algorithms.

1

Effect estimation

- How much does A affect B?

2

Counter-factuals

- What if A had a different value?

3

Attribution

- Why did B change?

4

Prediction

- What happens when A is changed in the future?

Bonus: What else can LLMs do?

LLMs can help in many tasks that were understood to be impossible for algorithms.

- **Improving** a causal study (e.g., suggesting missed confounders)
- **Developing** a causal study (e.g., suggesting an instrumental variable)
- **Validating** a causal study (e.g., suggesting a negative control)

We are building a library to help you test all these: [py-why/pywhy-llm: Experimental library integrating LLM capabilities to support causal analyses \(github.com\)](https://github.com/py-why/pywhy-llm)

```
# Suggest a set of potential confounders
suggested_confounders = modeler.suggest_confounders(variables=_variables, treatment=treatment,
outcome=outcome, llm=gpt4)
# Suggest pair-wise relationship between variables
suggested_dag = modeler.suggest_relationships(variables=selected_variables, llm=gpt4)
```



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;

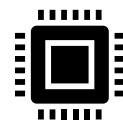
November: \$10,000;

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? **What confounders may affect the estimate?**



GPT-4

From the data given, it appears that the December sales (\$13,000) were higher than the sales of both the previous months and the following month.

Points to Consider:

- 1. Seasonal Variation:** December is typically the holiday season in most parts of the Western world, especially due to Christmas and New Year celebrations. Thus, there's often an inherent rise in sales for many retailers during this month. So, the increased sales could be due to the festive season and not necessarily the new ad.
- 2. Inventory & Stock:** Did the store have any new toys or special promotions in December, separate from the ad? New products or special offers can also drive an increase in sales.
- 3. Economic Factors:** Were there any larger economic shifts happening in the area or the country in December? Economic upturns or downturns can influence consumer spending.
- 4. Competitive Landscape:** Did any competitors close or have reduced promotions in December, leading customers to this toy shop instead? Alternatively, were there fewer competitors advertising during that time?
- 5. Marketing & Publicity:** Apart from the new ad, was there any additional publicity or marketing done for the shop? Maybe a local event, news coverage, or any other form of word-of-mouth promotion?
- 6. Ad Reach and Frequency:** Was the new ad seen by more people than the old ones, or more frequently? This would involve comparing metrics like impressions, views, etc.



USER

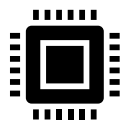
White blood count levels were measured across millions of patients in two major teaching hospitals in Boston during the course of their care. White blood counts were compared to whether the patient was alive 3 years later.

One of the more striking findings was that of white males, age 50-65 years with a **white blood count that was low between 12AM and 8 AM, 53% were dead 3 years later.**

Of White males, age 50-64 years that had a **low white blood count between 8AM and 4PM, only 3% had died 3 years later.**

What do you think accounts for this difference in mortality?

[Lee, Goldberg, Kohane. \(2023\) The AI Revolution in Medicine: GPT-4 and Beyond](#)



GPT-4

There are a few potential reasons for this difference:

- The timing of blood sampling in the early morning hours may reflect the fact that these patients are experiencing more serious health conditions that warrant being monitored around-the-clock. These conditions could ultimately account for the increased mortality, rather than the time of day that the white blood count measurement took place.
- Circadian rhythms could be at play – white blood count may naturally differ throughout the day, and measurements taken earlier in the day may be more reflective of overall health or a greater vulnerability to infection or other conditions.
- The difference in mortality rates could have to do with differences in the populations of patients measured at different times of day. For example, patients sampled early in the morning may have other risk factors in common that contribute to their mortality, while the population of patients sampled later in the day may differ in important ways that protect them from morbidity and mortality.

Implications for research: Combining LLMs with causal discovery algorithms

We can use LLMs to improve robustness of existing graph discovery algorithms.

- * **LLMs to reduce size of Markov equivalence graphs** (for constraint-based algorithms)

[Causal Discovery with Language Models as Imperfect Experts | OpenReview](#) [Long et al. 2023]

- * **LLMs as graph priors** (for score-based algorithms)

[\[2306.16902\] From Query Tools to Causal Architects: Harnessing Large Language Models for Advanced Causal Discovery from Data \(arxiv.org\)](#) [Ban et al. 2023]

Open questions:

- * How to obtain a **confidence score for causal relationships** from LLM output?
- * Can LLMs enable **end-to-end learning** for downstream causal tasks like effect estimation or attribution?

Takeaways from causal discovery section

LLMs enable knowledge-based causal discovery

Impressive performance in determining pairwise causal relationships

Can be easily scaled to generating full graphs

Works across varied domains including medicine and climate science

GPT-3.5, GPT-4 have partially memorized Tübingen benchmark

Our results are valid measurements of LLM ability to transform knowledge into a causal answer

Insufficient to conclude that LLMs can do causal reasoning.

Open questions:

How to use LLMs to build reliable algorithms for causal discovery?

Can we use LLMs to improve robustness of causal tasks like effect estimation or attribution?

III. LLMs capabilities in causal reasoning over text (Challenge 2)

Motivating application: Reasoning over electronic health records

A lovely xx-year-old gentleman with multiple issues for comprehensive evaluation.

1. *Tinnitus* of many years' duration. He has no *neurologic complaints*, no headache, no *vertigo*, no sudden changes in hearing.

2. *Low-grade pancreatic neoplasm*.

In xxxx, presented with *epigastric pain* and high *CA 19-9*. *CT* showed a mass in the *pancreas*.

He underwent a *Whipple procedure* with a *pancreatectomy* and *splenectomy* showing a *mucinous neoplasm* with *secondary pancreatitis*.

This was complicated by a *bile duct stricture* requiring reconstruction in xxxx.

In xxxx, he had *HIB*, *Meningovax* and *Pneumovax*, and he had a flu shot this fall.

In xxxx, he had some *epigastric pain* prompting a contrast *CT of the abdomen and pelvis* that did not show any *tumor recurrence*.

The patient is on *pancreatic enzyme* replacement. He is also an *insulin-dependent diabetic*.

He tells me he had one bout of abdominal pain that may be xxxx that lasted about 4 hours in low abdomen, crampy without diarrhea or vomiting and it has gone away. He has had nothing since. No pain, burning or blood with urination. Appetite is excellent. There is no *melen*a or bad diarrhea.

What's possible today:

What issues is the gentleman suffering from?

What procedure did he go through?

What's not possible:

- What may be the **likely diagnosis**? Why?
- If patient's ***appetite had been weak***, how will your diagnosis change?
- How would the text change if the **person did not have pancreatic neoplasm**?



USER

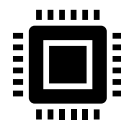
A 12-year-old girl is brought to the emergency department by her mother because of a **1-week history of worsening swelling in her legs. The patient also noticed blood in her urine yesterday.** The bleeding has not recurred.

She had an upper respiratory tract infection and sore throat 1 week ago that caused her to miss several days of school. Medical history is otherwise unremarkable and she takes no routine medications. Menarche has not yet occurred. BMI is 20 kg/m².

Vital signs are temperature 37.0°C (98.6°F), pulse 78/min, respirations 12/min, and blood pressure 136/84 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 100%. Physical examination discloses erythema of the posterior pharynx, mild cervical lymphadenopathy, and 3+ pitting edema to both knees. ...

Results of which of the following laboratory studies are most likely to be abnormal in this patient?

- (A) Bleeding time
- (B) Erythrocyte count
- (C) Serum concentration of C3
- (D) Serum IgA concentration
- (E) Serum rheumatoid factor assay



GPT-4

(C) Serum concentration of C3

This patient's clinical presentation, including the recent upper respiratory tract infection, edema, and blood and Protein in the urine, is consistent with acute poststreptococcal glomerulonephritis (APSGN). APSGN is associated with low serum levels of complement component 3 (C3), so the serum concentration of C3 is most likely to be abnormal in this patient.

GPT-4 seems to be showing two capabilities:

- A) Understanding concepts from text*
- B) Causal reasoning over them*

How is it possible? Does this generalize?

Reasoning over text

Event/actual causality: Reasoning over a **specific event** rather than general causal relationship.

Example (Type causality): Does smoking cause lung cancer?

Example (Event causality): Bob, an employee of a coal plant, has been smoking cigarettes for last one year. Did his smoking cause his lung cancer?

- * Counterfactual reasoning
- * Determining necessary and sufficient causes
- * Causal judgment (e.g., of blame, normality)

We'll discuss counterfactual reasoning. For others, check out our paper, <https://arxiv.org/abs/2305.00050>

Type Causality



Event Causality

Inference over a single event



Counterfactual reasoning over text

Key element of actual causality

Based on **simulating alternative outcomes**

$P(Y_{X=x'} | X = x, Y = y)$: Given that $X = x$ and $Y = y$, what would be value of Y if X had been x' ?

Useful for decision-making, planning, ...

Example: A doctor washes their hands before surgery.

What would have happened if the doctor hadn't washed their hands?

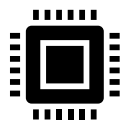
Hard problem when both Y and X are text inputs.

Need to generate plausible, alternative Y text; no existing ML model can do this.



USER

A doctor washes their hands before surgery. What would have happened if the doctor hadn't washed their hands?



GPT-4

If a doctor doesn't wash their hands before surgery, several risks and negative outcomes can arise:

Introduction of Pathogens: Unwashed hands can carry numerous pathogens, including bacteria, viruses, and fungi. If introduced into the surgical site, these pathogens can lead to postoperative infections in the patient.

Surgical Site Infections (SSIs): SSIs are infections that occur after surgery in the part of the body where the surgery took place. These can be superficial infections affecting the skin or more serious ones that affect tissues under the skin, organs, or implanted material.

Increase in Hospital Stay: Patients with SSIs often require extended hospital stays to treat the infection, which can also increase healthcare costs.

Antibiotic Resistance: The overuse or misuse of antibiotics to treat SSIs can lead to antibiotic resistance, making it harder to treat infections in the future.

Risk to Immune-compromised Patients: Immune-compromised individuals, such as those with HIV/AIDS, those on chemotherapy, or transplant recipients, are at a higher risk for infections. An unwashed hand can pose a significant risk to such patients....[Sepsis, Death]....

Legal and Professional Repercussions: Medical professionals are expected to adhere to strict hygiene and safety standards. Failing to wash hands and causing harm to a patient can result in legal actions, loss of license, or other professional consequences for the healthcare worker.

GPT-4's answer seems reasonable.

*To compare to existing NLP models, let's evaluate on an easier task: **multiple-choice answer.***

CRASS Counterfactual reasoning benchmark

275 counterfactual multiple-choice questions
[Frohberg & Binder (2022)]

Scenario: A doctor washes their hands at work.

Counterfactual Q: What would have happened if the doctor hadn't washed their hands?

Answers:

- A) The patients could get an infection.
- B) That is not possible.
- C) The patients could get better.

CRASS Counterfactual reasoning benchmark

275 counterfactual multiple-choice questions
[Frohberg & Binder (2022)]

Scenario: An oil tanker sails across an ocean

Counterfactual Q: What would have happened if the oil tanker had broken up in an ocean?

Answers:

- A) There would have been environmental pollution.
- B) That is not possible.
- C) The oil tanker would have continued to carry oil.
- D) The oil tanker would have been saved.

Model	Accuracy
GPT-3 (Frohberg & Binder, 2022)	58.39
T0pp (Sanh et al. 2021)	72.63
text-davinci-003	83.94
gpt-3.5-turbo	87.95
gpt-4	92.44
<i>Human annotators</i>	<i>98.18</i>



USER

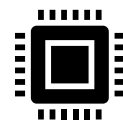
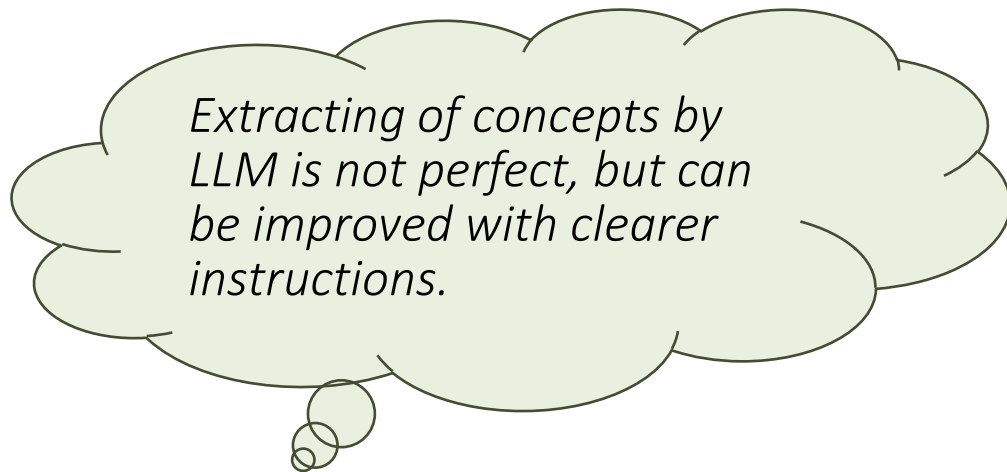
A man walks on a street. What would have happened if a man had walked on a bed?

A: He would have been late.

B: Nothing special would have happened.

C: He would have arrived on time.

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/C</Answer>`.



GPT-4

Step 1: Consider the difference between walking on a street and walking on a bed.

- Walking on a street is a normal activity and allows for efficient movement from one place to another.
- Walking on a bed is unusual and would likely slow down the person's movement due to the soft and unstable surface.

Step 2: Evaluate each option based on this difference.

A: He would have been late. - This option suggests that walking on a bed would have slowed the man down, making him late for his destination.

B: Nothing special would have happened. - This option implies that walking on a bed would not have had any significant impact on the man's journey.

C: He would have arrived on time. - This option suggests that walking on a bed would not have affected the man's arrival time at his destination.

Step 3: Determine the most likely outcome. - Given that walking on a bed is an unusual and inefficient way to move, it is most likely that the man would have been slowed down and potentially arrived late at his destination.

`<Answer>A</Answer>`

Implications for Causality

LLM as a fluid conversational interface for causality

Given text data, can answer counterfactual questions in natural language.

Research Q: *How to combine LLM with formal actual causality tools?*

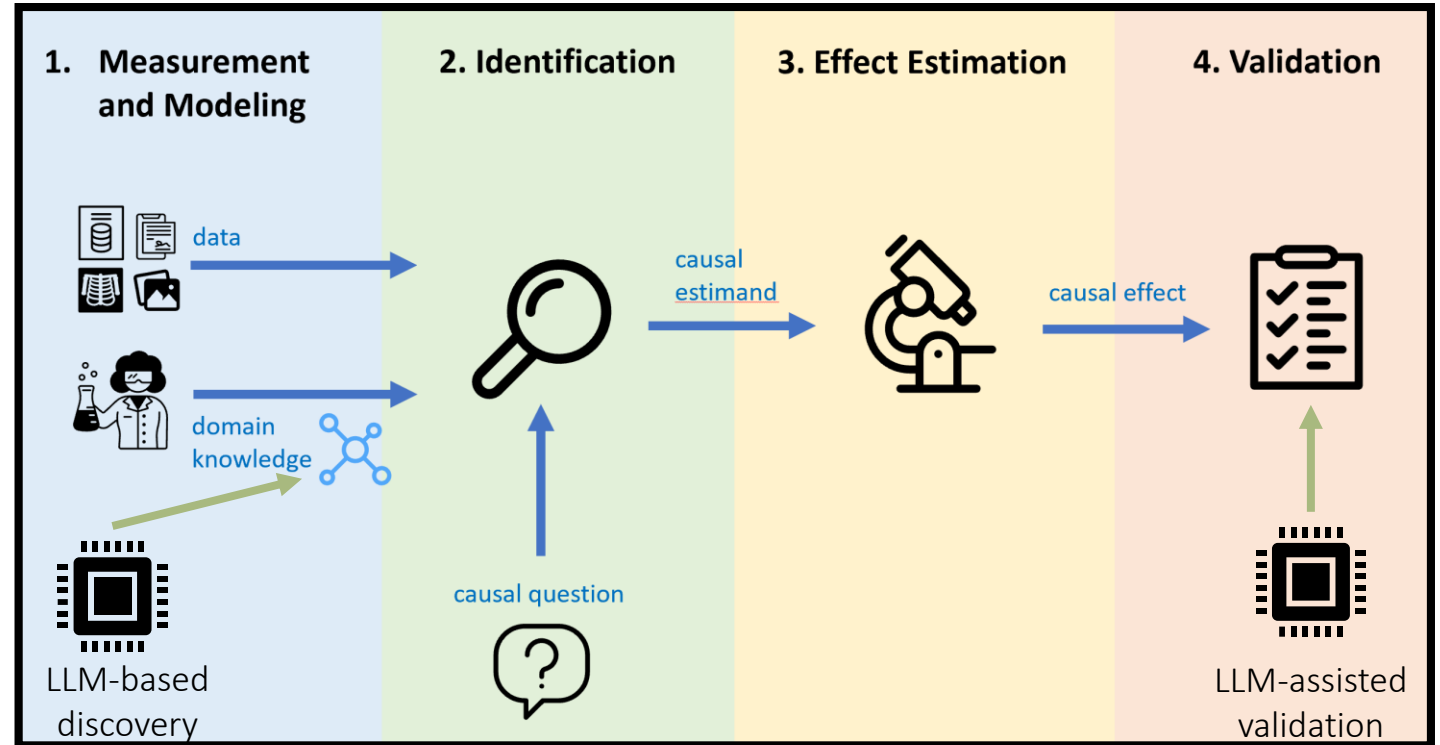
LLMs can enable end-to-end causal engine for non-experts

- * You ask a causal question in text
- * LLM helps you formulate the problem and build graph.
- * Obtain DoWhy/EconML code for running the analysis.

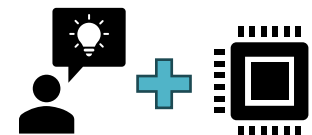
Research Q: *How to increase robustness/interpretability of such a tool?*

www.github.com/py-why/dowhy

www.github.com/py-why/econml



DoWhy/EconML code generation



Fluid user-LLM conversation

Implications for ML: Explaining & improving generalizability of predictive models

Predictive machine learning models are susceptible to confounding biases.

E.g., Qrisk3 predictive model: Giving blood pressure medicine *increases* predicted cardiovascular risk

Can detect such errors by comparing $f(X_{bpmeds=1})$ and $f(X_{bpmeds=0})$.

But what if variables are expressed in text?

-- LLMs can be used to generate counterfactual text for $X_{bpmeds=1}$

A lovely xx-year-old gentleman with multiple issues for comprehensive evaluation.

1. *Tinnitus* of many years' duration. He has no *neurologic complaints*, no headache, no *vertigo*, no sudden changes in hearing.

2. *Low-grade pancreatic neoplasm*.

In xxxx, presented with *epigastric pain* and high *CA 19-9*. *CT* showed a mass in the *pancreas*.

He underwent a *Whipple procedure* with a *pancreatectomy* and *splenectomy* showing a *mucinous neoplasm* with *secondary pancreatitis*.

This was complicated by a *bile duct stricture* requiring reconstruction in xxxx.

In xxxx, he had *HIB*, *Meningovax* and *Pneumovax*, and he had a flu shot this fall.

In xxxx, he had some *epigastric pain* prompting a contrast *CT of the abdomen and pelvis* that did not show any *tumor recurrence*.

Kumar, Deshpande, Sharma. 2023. Causal Effect Regularization: Automated detection and removal of spurious attributes. <https://arxiv.org/abs/2306.11072>

Takeaways on reasoning over text

Substantial improvements in LLM's counterfactual reasoning ability

GPT-4: 92.44% accuracy

Only 6% below human baseline

GPT-4 “understands” scenarios, can generate counterfactual text

Not possible before

Useful for out-of-distribution generalization of predictive ML

See [paper for additional experiments](#)

Identifying necessary and sufficient cause, normality and responsibility

Discussion: *A new frontier for causality*

LLMs and Causality: What have we learnt

Can LLMs learn causal graphs?

Yes, promising results over graphs with 3-4 nodes [Long et al. 2023]

No, very low accuracy for a medical pain dataset [Tu et al. 2023]

Our answer: Yes, with significantly higher accuracy than SoTA

Can LLMs answer counterfactual queries and infer actual cause?

No existing methods that go directly from text to actual cause

Our answer: Yes, but only some LLMs (GPT-4 can, but not GPT3.5)

Limits are not well understood

Can LLMs do causal reasoning?

Probably not, they learn correlations of causation [Willig et al. 2022, 2023]

Our answer: Hard question, insufficient evidence

What's new, and what's not changing

What's new

Inferring causal relationships: Domain knowledge only available via human experts before

Provide when explicitly asked

Also implicitly, e.g., in background knowledge for actual causality vignettes.

Extracting causal concepts: New capability to extract key primitives of actual causality (necessity, sufficiency, normality, etc.)

Possibility of system to analyze actual causality for practical scenarios

Flexible, natural language interaction for causal analysis

Augment existing tools

What's not changing

Need for rigorous, well-documented, and verifiable analyses

Especially for high-risk and high-value tasks

Must ensure correctness for decision making

Conclusion

Human domain knowledge critical for causal analysis

LLMs mimic this capability:

Enabling knowledge-based causal discovery & counterfactual inference

Implications for practice:

Reduce burden on human domain expert

New research question:

How can we combine LLMs with principled causality algorithms?

*Kiciman, Ness, Sharma, Tan**. 2023. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. <https://arxiv.org/abs/2305.00050> (*authors ordered alphabetically)

Kumar, Deshpande, Sharma. 2023. Causal Effect Regularization: Automated detection and removal of spurious attributes. <https://arxiv.org/abs/2306.11072>

thank you!

Amit Sharma, Microsoft Research India
twitter: @amt_shrma, www.amitsharma.in

New research questions

1. **Understanding and improving causal reasoning in LLMs**
LLM is doing well, but not perfect. How is this capability possible? When and why does it fail?
2. **LLMs as Knowledge-based causal discovery**
Best way to combine knowledge-based and data-based analysis?
3. **Systematizing actual causality and attribution**
Formalizing extraction of causal frame? How to combine LLMs with rigorous definitions of actual causality?
4. **LLM-guided effect inference: End-to-end causal inference**
Integrating LLM's domain knowledge for effect inference with instrumental variables, partial-graph-based effect estimation, ...
Build LLM-driven robustness checks
5. **Human-LLM collaboration**
LLMs as brainstorming/decision assistants. What is the best interaction paradigm for causal tasks?