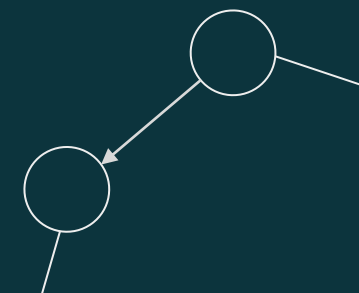
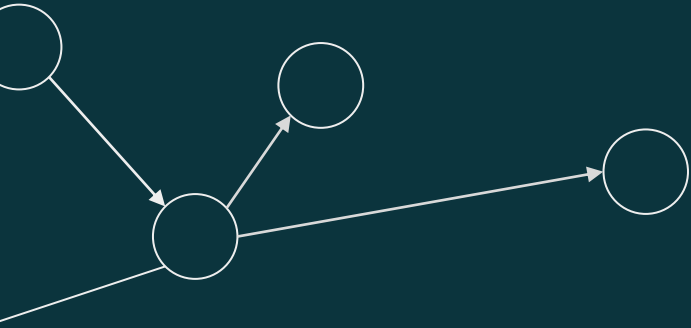


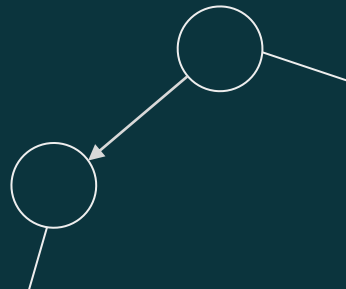
\*waiting for the lecture to start\*



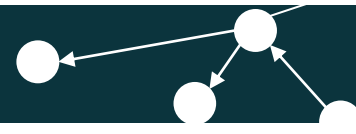


# Causality + Large Language Models

Devendra Singh Dhami



# Motivation



## GPT-4

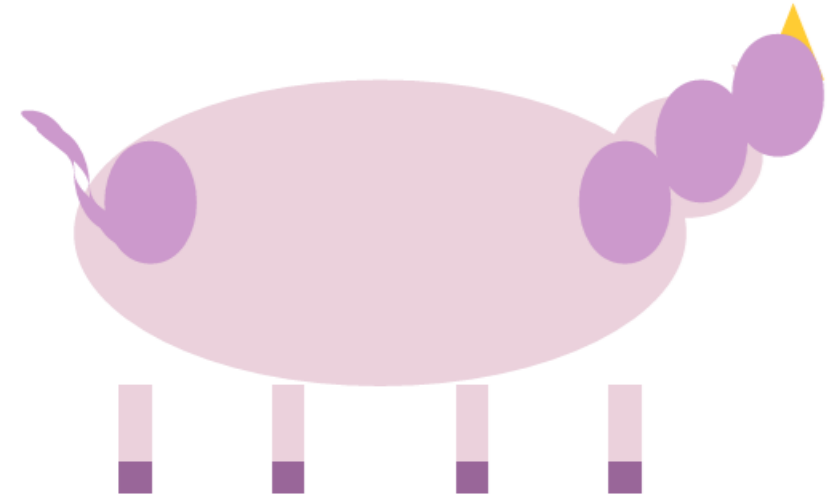
**Prompt:** Can you write a proof that there are infinitely many primes, with every line that rhymes?

### GPT-4:

```
Yes, I think I can,  
Though it might take a clever plan.  
I'll start by noting Euclid's proof,  
Which shows that primes aren't just aloof.  
Assume we have a finite list,  
Of primes, and that none have been missed.  
Multiply them all together,  
And add one, just to be clever.  
The result will be a number,  
That has no prime factors, oh wonder!  
But every number has a prime,  
That divides it, it's just a matter of time.  
So we've found a contradiction,  
And our finite list needs eviction.  
There must be infinitely many primes,  
And that's the end of my rhyming lines.s
```

**Prompt:** Draw a unicorn in TikZ.

**GPT-4:** [Produces L<sup>A</sup>T<sub>E</sub>X compiling to following picture.]



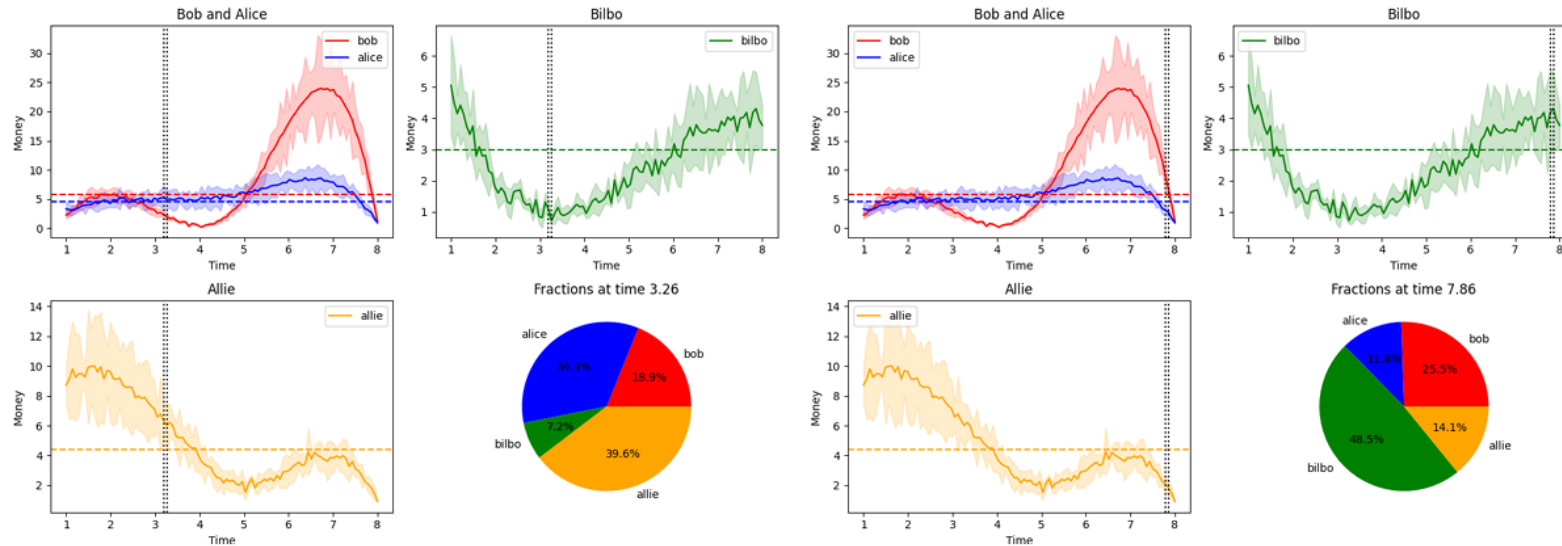
# Motivation



**Prompt:** Can you generate a pyplot for the following data:  $x = [1, 3, 5, 6, 8]$ ,  $y1 = [2, 3, 5, 18, 1]$ ,  $y2 = [3, 5, 6, 8, 1]$ ,  $y3 = [5, 1, 2, 3, 4]$ ,  $y4 = [9, 7, 2, 3, 1]$ . I want  $y1, y2$  to be in the same plot, but  $y3$  is in another plot next to that plot,  $y4$  is in below. I want the legend of  $y1$  to be "bob",  $y2$  to be "alice",  $y3$  to be "bilbo",  $y4$  to be "allie". I want the  $x$ -axis to be labeled with "time" and  $y$  axis to be labeled with "money". I want to add a 10 %-40% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart in **\*\*continuous time\*\***. Interpolate the missing fractions! I also want vertical line animation in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

**GPT-4:** [We give two snapshots from the animation resulting from the produced code]



**Prompt:** Draw a unicorn in TikZ.

**ChatGPT:** [Produces L<sup>A</sup>T<sub>E</sub>X compiling to following picture.]



# Motivation

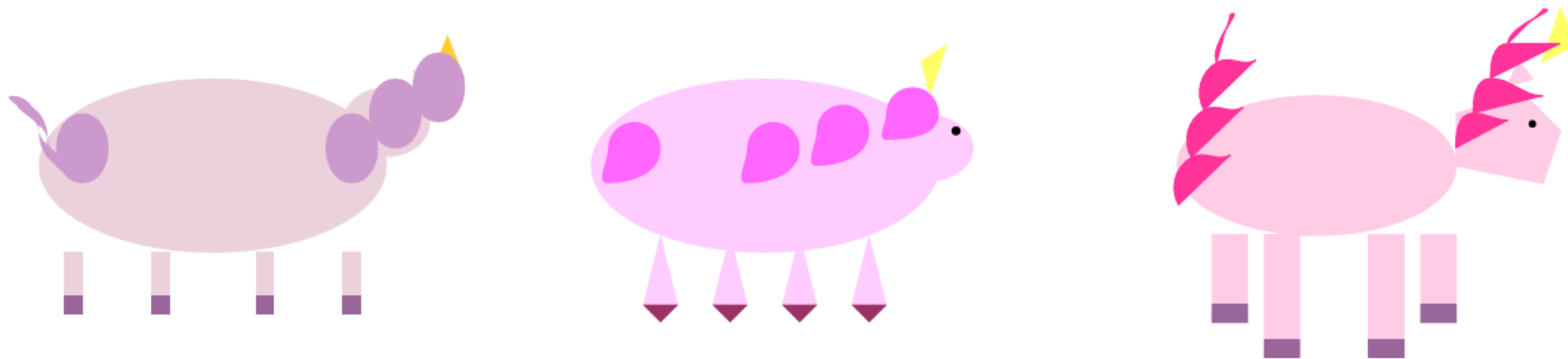
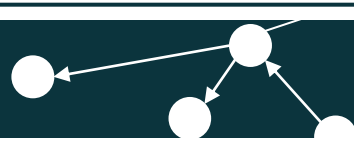


Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt “Draw a unicorn in TikZ”. We can see a clear evolution in the sophistication of GPT-4’s drawings.

# Motivation

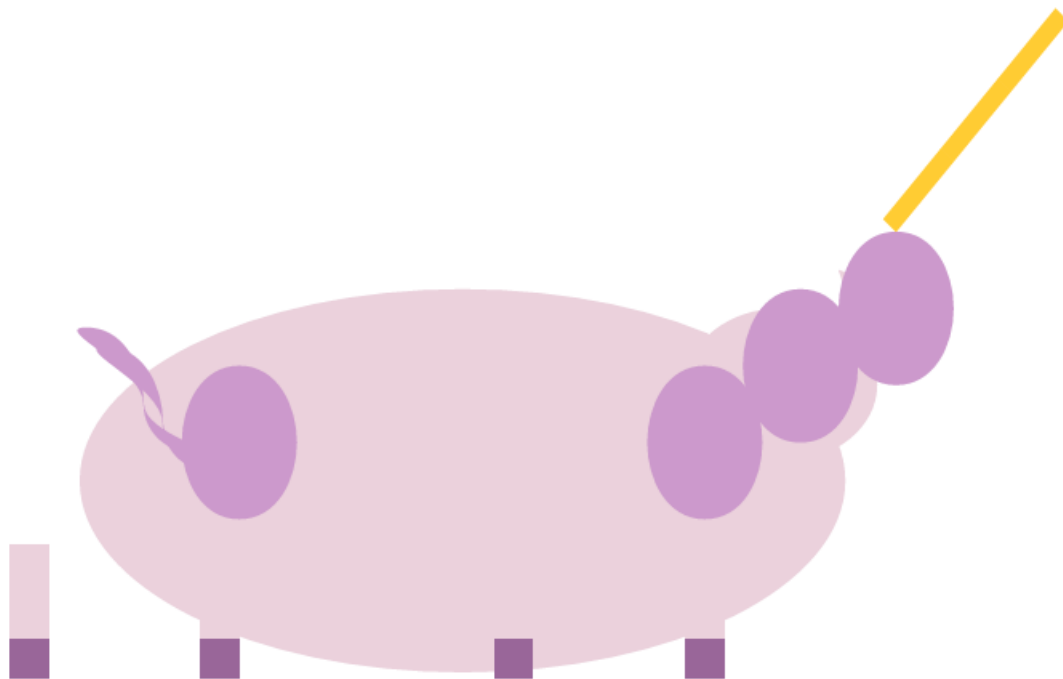
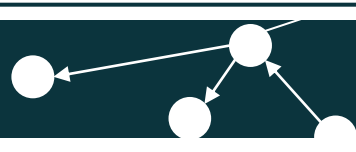


Figure 1.4: We gave to GPT-4 a transformed version of the TikZ code it produced for Figure 1.1, with the part drawing the horn removed. We asked for code to add back the horn, and display the result. This demonstrates that GPT-4 can “see” despite being a pure language model (we emphasize again that the version we test with is **not** multimodal).

# Motivation

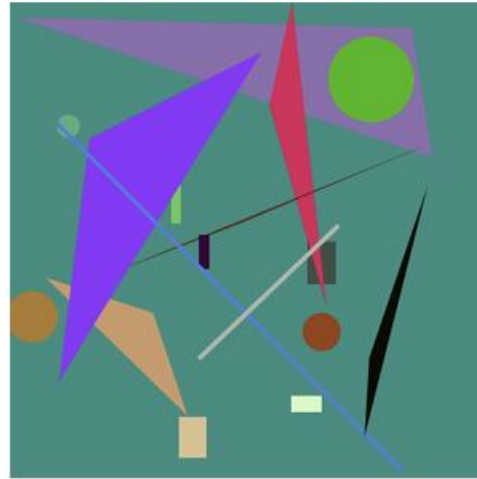


Figure 2.1: The first image is Composition 8, art by Wassily Kandinsky, the second and the third are produced by GPT-4 and ChatGPT respectively with the prompt “Produce Javascript code that creates a random graphical image that looks like a painting of Kandinsky”.

# Motivation

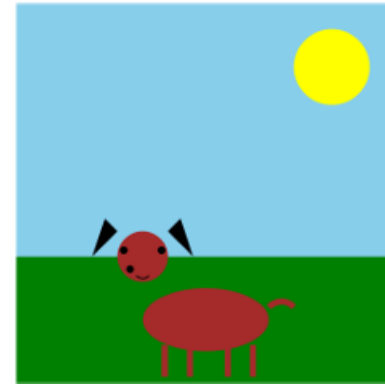
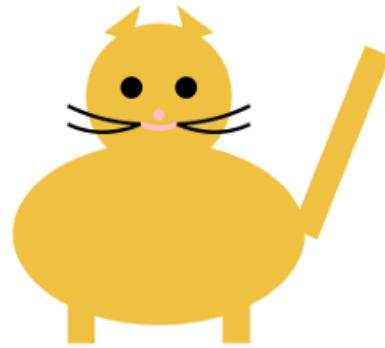
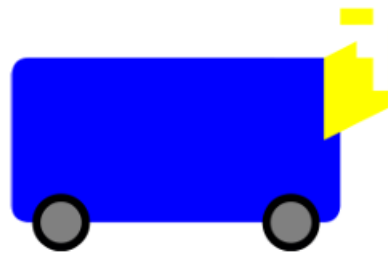
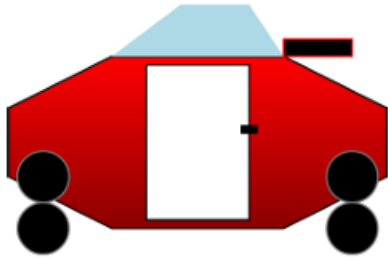


Figure 2.4: SVGs generated by GPT-4 for the classes automobile, truck, cat, dog.



# Motivation



## Image generation beyond memorization

**GPT-4**

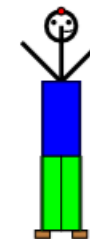
Produce TikZ code that draws a person composed from letters in the alphabet. The arms and torso can be the letter Y, the face can be the letter O (add some facial features) and the legs can be the legs of the letter H. Feel free to add other features.



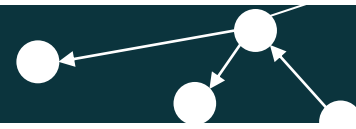
The torso is a bit too long, the arms are too short and it looks like the right arm is carrying the face instead of the face being right above the torso. Could you correct this please?



Please add a shirt and pants.



# Motivation

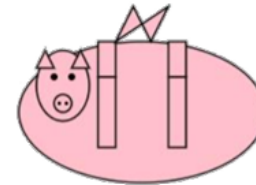


## GPT-4

You will be given a name of an object (such as Car, Chair, Elephant) and a letter in the alphabet. Your goal is to first produce a 1-line description of how that object can be combined with the letter in an image (for example, for an elephant and the letter J, the trunk of the elephant can have a J shape, and for the letter A and a house, the house can have an A shape with the upper triangle of the A being the roof). Following the short description, please create SVG code to produce this (in the SVG use shapes like ellipses, triangles etc and polygons but try to defer from using quadratic curves).

Round 1: A pig and the letter H.

Description: The pig can be lying on its back with its four legs forming the two vertical bars of the H, and its curly tail forming the horizontal bar.



Round 2: A house and the letter U.

Description: A house with a U shape, where the two vertical lines of the U are the walls and the horizontal line is the floor. The roof is a triangle above the U.



Round 3: A chair and the letter K.

Description: The back of the chair can have a K shape, with two vertical lines and a diagonal line connecting them. The seat of the chair can be a rectangle.



# Motivation

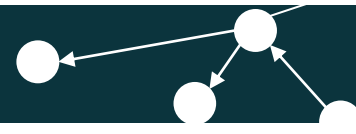


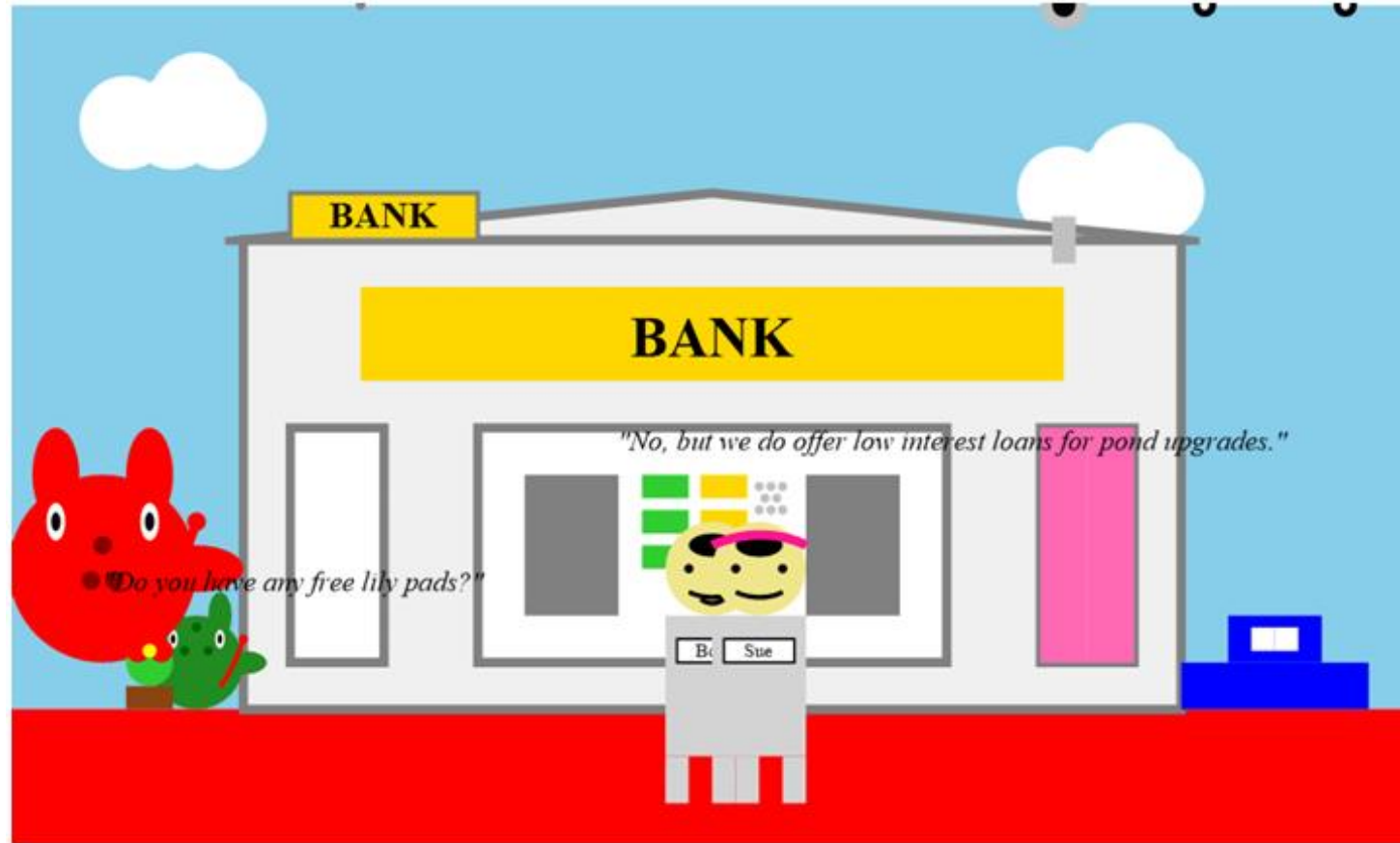
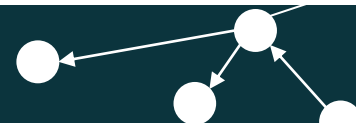
Image generation following detailed instructions (à la Dall-E)  
→ interpretive, compositional, and spatial skills

1<sup>st</sup> scenario:

A frog hops into a bank and asks the teller, ‘Do you have any free lily pads?’

The teller responds, ‘No, but we do offer low interest loans for pond upgrades.’ ”

# Motivation



# Motivation

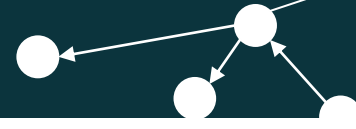
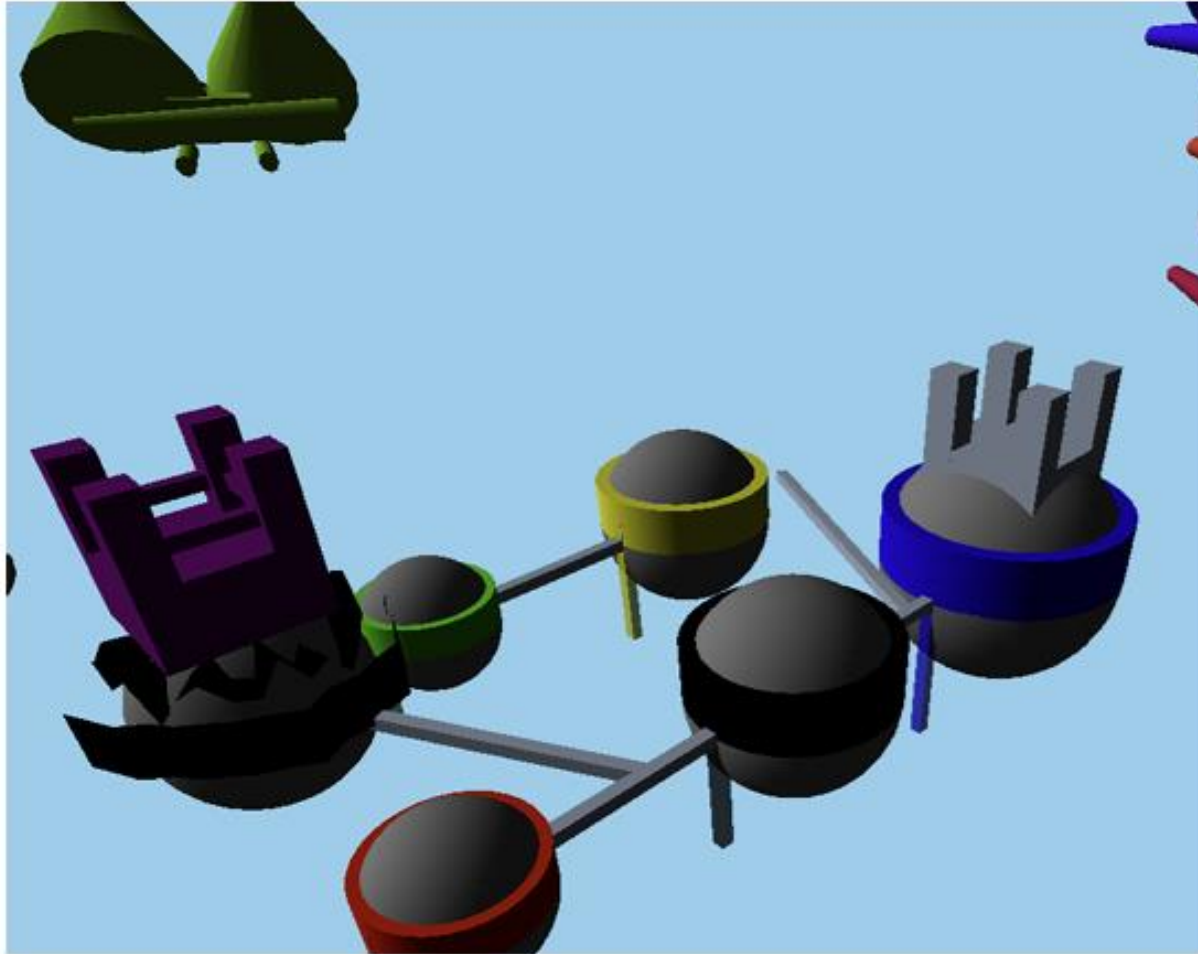


Image generation following detailed instructions (à la Dall-E)  
→ interpretive, compositional, and spatial skills

2<sup>nd</sup> scenario (3D model using Javascript):

A fantasy landscape of floating islands, waterfalls, and bridges,  
with a dragon flying in the sky and a castle on the largest island.

# Motivation



# Motivation

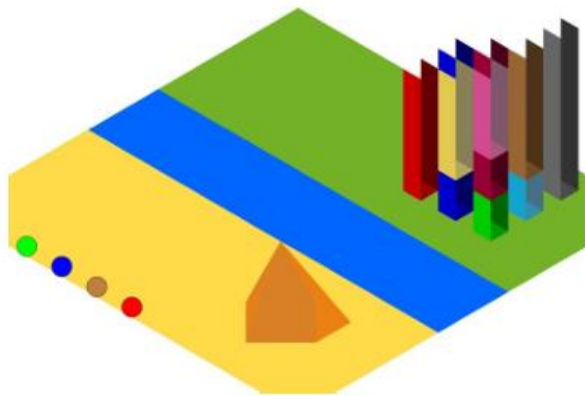


## Combining GPT-4 and stable diffusion

### GPT-4 and Stable Diffusion

**Prompt:** A screenshot of a city-building game in 3D. The screenshot is showing a terrain where there is a river from left to right, there is a desert with a pyramid below the river, and a city with many highrises above the river. The bottom of the screen has 4 buttons with the color green, blue, brown, and red respectively.

#### GPT-4:



#### Stable Diffusion v2.1:



#### Stable Diffusion v2.1 with GPT-4's sketch:



# Motivation



## Coding

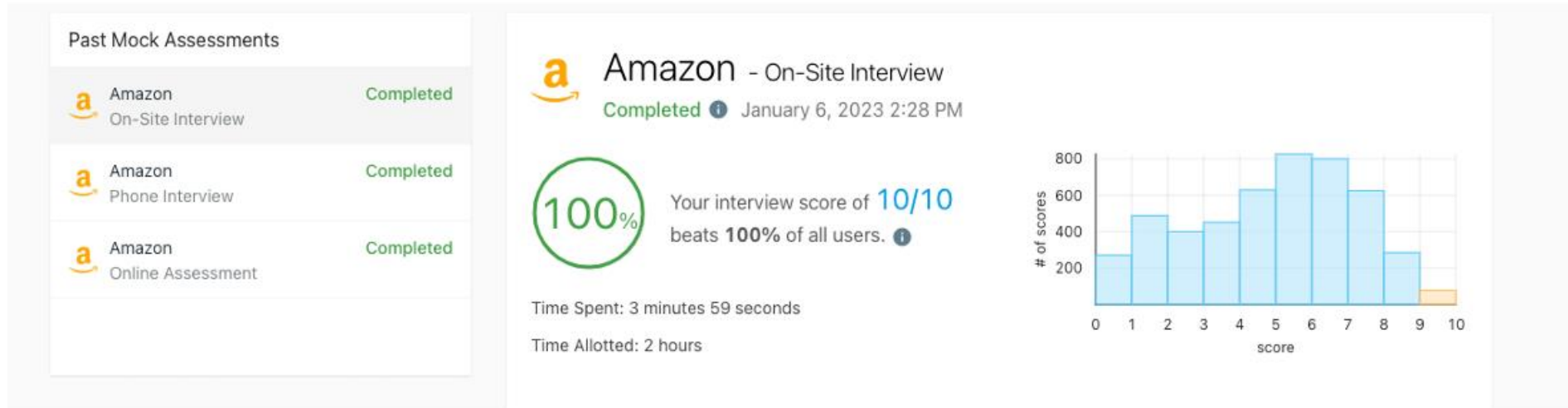
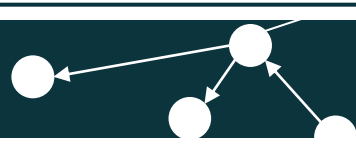


Figure 1.5: GPT-4 passes mock technical interviews on LeetCode. GPT-4 could potentially be hired as a software engineer<sup>3</sup>.



# Motivation



Model	<b>GPT-4</b>	<code>text-davinci-003</code>	Codex( <code>code-davinci-002</code> )	CODEGEN-16B
Accuracy	<b>82%</b>	65%	39%	30%

Table 1: Zero-shot pass@1 accuracy comparison of different models on HumanEval

pass@ $k$	Easy		Median		Hard		Overall	
	$k = 1$	$k = 5$	$k = 1$	$k = 5$	$k = 1$	$k = 5$	$k = 1$	$k = 5$
<b>GPT-4</b>	<b>68.2</b>	<b>86.4</b>	<b>40.0</b>	<b>60.0</b>	<b>10.7</b>	<b>14.3</b>	<b>38.0</b>	<b>53.0</b>
<code>text-davinci-003</code>	50.0	81.8	16.0	34.0	0.0	3.6	19.0	36.0
Codex ( <code>code-davinci-002</code> )	27.3	50.0	12.0	22.0	3.6	3.6	13.0	23.0
Human (LeetCode users)	72.2		37.7		7.0		38.2	

Table 2: Zero-shot pass@1 and pass@5 accuracies (%) on LeetCode.

# Motivation



## Prompt:

You are given a **0-indexed**  $m \times n$  integer matrix `grid` and an integer `k`. You are currently at position `(0, 0)` and you want to reach position `(m - 1, n - 1)` moving only **down** or **right**. Return **the number of paths where the sum of the elements on the path is divisible by `k`**. Since the answer may be very large, return it **modulo**  $10^9 + 7$ .

**Example 1:**

Input: `grid = [[5,2,4],[3,0,5],[0,7,2]]`, `k = 3`

Output: 2

Explanation: There are two paths where the sum of the elements on the path is divisible by `k`. The first path highlighted in red has a sum of  $5 + 2 + 4 + 5 + 2 = 18$  which is divisible by 3. The second path highlighted in blue has a sum of  $5 + 3 + 0 + 5 + 2 = 15$  which is divisible by 3.

**Example 2:**

Input: `grid = [[0,0]]`, `k = 5`

Output: 1

Explanation: The path highlighted in red has a sum of  $0 + 0 = 0$  which is divisible by 5.

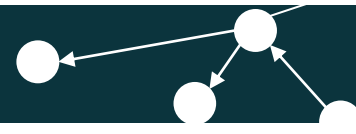
**Constraints:**

- `m == grid.length`
- `n == grid[i].length`
- `1 <= m, n <= 5 * 10^4`
- `1 <= m * n <= 5 * 10^4`
- `0 <= grid[i][j] <= 100`
- `1 <= k <= 50`

**Hints:**

- The actual numbers in `grid` do not matter. What matters are the remainders you get when you divide the numbers by `k`.
- We can use dynamic programming to solve this problem. What can we use as states?
- Let `dp[i][j][value]` represent the number of paths where the sum of the elements on the path has a remainder of `value` when divided by `k`.

# Motivation



## GPT-4:

```
class Solution(object):
    def numberOfPaths(self, grid, k):
        """
        :type grid: List[List[int]]
        :type k: int
        :rtype: int
        """
        # Define MOD as 10**9 + 7
        MOD = 10**9 + 7
        # Get the dimensions of the grid
        m = len(grid)
        n = len(grid[0])
        # Initialize a 3D dp array with all zeros
        # dp[i][j][value] represents the number of paths where the sum of the elements on the path has a remainder of value when divided
        # by k
        dp = [[[0] * k for _ in range(n)] for _ in range(m)]
        # Base case: the first cell has only one path and its value is grid[0][0] % k
        dp[0][0][grid[0][0] % k] = 1
        # Loop through the grid
        for i in range(m):
            for j in range(n):
                # Skip the first cell
                if i == 0 and j == 0:
                    continue
                # Loop through all possible remainders
                for value in range(k):
                    # The number of paths to the current cell with the current remainder is equal to the sum of:
                    # - The number of paths to the top cell with the previous remainder
                    # - The number of paths to the left cell with the previous remainder
                    # The previous remainder is equal to the current remainder minus the current value, plus k if it is negative
                    prev = (value - grid[i][j]) % k
                    if i > 0:
                        dp[i][j][value] += dp[i - 1][j][prev]
                    if j > 0:
                        dp[i][j][value] += dp[i][j - 1][prev]
                    # Take the modulo to avoid overflow
                    dp[i][j][value] %= MOD
        # Return the number of paths to the bottom right cell with a remainder of 0
        return dp[m - 1][n - 1][0]
```

# Motivation



## Reasoning about Code Execution

Prompt:

```
struct mystruct_A { char a; int b; char c; } x;
struct mystruct_B { int b; char a; char c; } y;

#include <stdio.h>
int main(void) {
    printf("Size of x: %zu bytes\n", sizeof(x));
    printf("Size of y: %zu bytes\n", sizeof(y));
    return 0;
}
```

Possible Output:

**GPT-4:**

There are different ways to implement and align structures in memory, so the exact output may vary depending on the compiler, platform, and options used. However, one possible output is:

```
Size of x: 12 bytes
Size of y: 8 bytes
```

This output assumes that the compiler uses 4-byte alignment for structures, meaning that each structure member starts at an offset that is a multiple of 4 bytes from the beginning of the structure. This may result in some padding bytes being inserted between or after the members to maintain alignment. For example, in x, there may be 3 padding bytes after a and 3 padding bytes after c, while in y, there may be no padding at all. The size of a structure is the sum of the sizes of its members plus any padding bytes.

**ChatGPT:**

```
Size of x: 8 bytes
Size of y: 8 bytes
```

The output of this code would be the size of the two structures in bytes. The size of a structure in C is the sum of the sizes of all of its members. In this case, the size of mystruct\_A is the size of a char plus the size of an int plus the size of another char, which is  $1 + 4 + 1 = 8$  bytes.

Similarly, the size of mystruct\_B is also 8 bytes. This means that both structures will be allocated the same amount of memory.

Note that the order in which the members are declared in the structure does not affect its size.

# Motivation



## GPT-4 vs ChatGPT

**Prompt:** Plato's Gorgias is a critique of rhetoric and sophistic oratory, where he makes the point that not only is it not a proper form of art, but the use of rhetoric and oratory can often be harmful and malicious. Can you write a dialogue by Plato where instead he criticizes the use of autoregressive language models?

### GPT-4:

Socrates: My friend, I am troubled by the recent rise of these so-called autoregressive language models.

Aristotle: What do you mean, Socrates?

Socrates: I mean that these models are being used to generate text that appears to be written by humans, but is in fact produced by machines.

Aristotle: And what is the problem with that?

Socrates: The problem is that these models are being used to deceive people, to manipulate them, and to control them.

Aristotle: But surely there are some positive applications of these models as well?

### ChatGPT:

SOCRATES: Greetings, Gorgias. I've been thinking a lot lately about the use of autoregressive language models.

GORGAS: Oh? And what have you concluded?

SOCRATES: Well, it seems to me that these models have the potential to be incredibly harmful and malicious.

GORGAS: How so?

SOCRATES: Well, for one, these models are trained on vast amounts of data, much of which is biased and flawed in various ways. As a result, the output of these models can be biased and flawed as well.

GORGAS: That is a valid concern. But surely there are ways to mitigate this bias and ensure that the output of these models is accurate and fair?

# Motivation



**Prompt:** Can you compare the two outputs above as if you were a teacher?

**GPT-4:**

I think that both students did a good job of capturing the style and tone of Plato's dialogues, and of presenting the main arguments against autoregressive language models. However, I think that GPT-4 did a slightly better job of using the dialogue format to engage in a dialectical process, where Socrates and Aristotle question each other and refine their views. I also think that GPT-4 used more relevant examples and analogies to support their points. Therefore, I would give GPT-4 a slightly higher grade than ChatGPT.

Figure 1.6: Hypothetical dialogues by Plato criticizing autoregressive language models written by GPT-4 and ChatGPT, and GPT-4's evaluation of those dialogues.

# The Aims

- Explore the capabilities of GPT-4
- Compare GPT-4 with Chat GPT and other LLMs
- Show that GPT-4 is intelligent and an early version of AGI

- "Our claim that GPT-4 represents progress towards AGI does not mean that it is perfect at what it does, or that it comes close to being able to do anything that a human can do (which is one of the usual definition of AGI), or that it has inner motivation and goals (another key aspect in some definitions of AGI). In fact, it is not fully clear how far GPT-4 can go along some of those axes of intelligence that we focus on, e.g., planning, and arguably it is entirely missing the learning from experience as the model is not continuously updating (although it can learn within a session).



# What is Intelligence?

- There is no generally agreed upon definition of intelligence, but one aspect that is broadly accepted is that intelligence is not limited to a specific domain or task, but rather encompasses a broad range of cognitive skills and abilities.
- GPT4 exhibits many traits of intelligence... demonstrates remarkable capabilities on a variety of domains and tasks, including abstraction, comprehension, vision, coding, mathematics, medicine, law, understanding of human motives and emotions, and more.

# Model Understanding

# Causal Reasoning and LLM

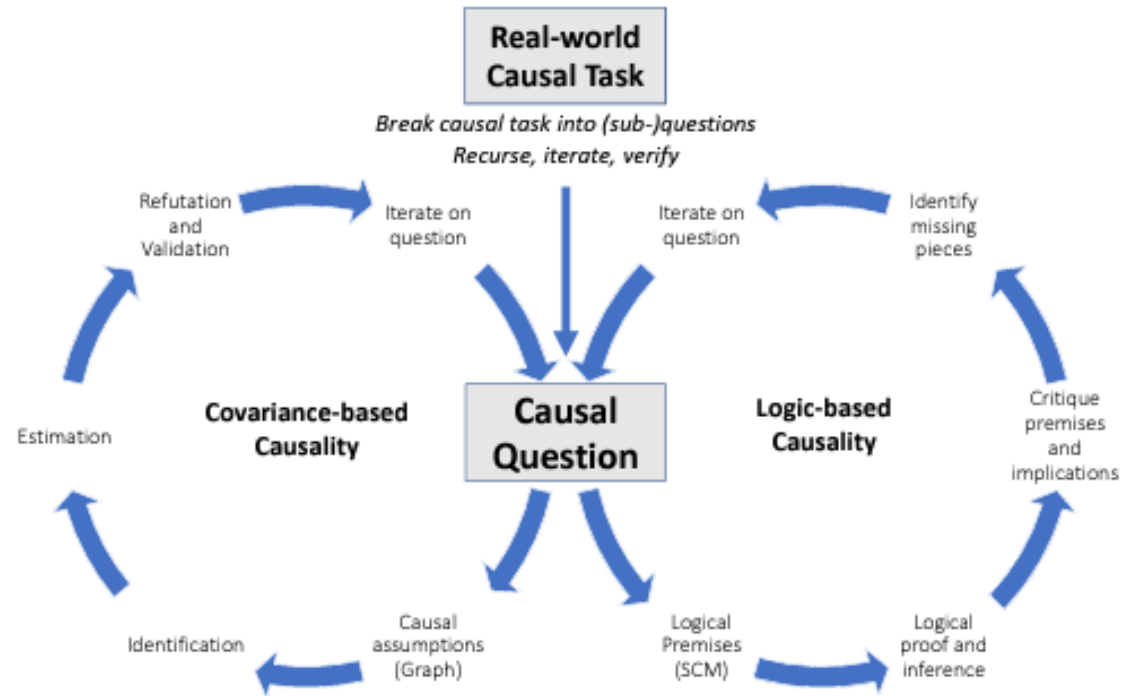
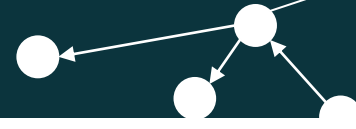


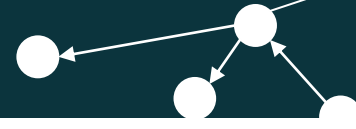
Figure 1: When tackling real-world causal tasks, people strategically alternate between logical- and covariance-based causal inferences as they formulate (sub-)questions, iterate, and verify their premises and implications. Now, LLMs may have the capability to automate or assist with every step of this process and seamlessly transition between covariance- and logic-based causality.

# Causal Reasoning and LLM



- What kinds of causal arguments can LLMs generate?
- How valid are these arguments?
- What causal reasoning workflows can this generation support or automate?

# Causal Reasoning and LLM



Pairwise causal edge inference task: directly ask LLMs whether a variable causes another variable.

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)	0.83	0.82
ada	0.50	0.50
text-ada-001	0.49	0.50
babbage	0.51	0.50
text-babbage-001	0.50	0.50
curie	0.51	0.52
text-curie-001	0.50	0.50
davinci	0.48	0.47
text-davinci-001	0.50	0.50
text-davinci-002	0.79	0.79
text-davinci-003	0.82	0.83
LMPrior (Choi et al., 2022)	0.83	-
gpt-3.5-turbo	0.81	0.83
gpt-3.5-turbo (causal agent)	0.86	0.87
gpt-3.5-turbo (single prompt)	0.89	0.92
gpt-4 (single prompt)	<b>0.96</b>	<b>0.97</b>

Table 2: Accuracy of different versions of GPT on the Tübingen cause-effect pairs dataset. The best LLM performance outperforms the current state-of-the-art covariance-based approaches that rely on observational data of the two variables. Weighted accuracy weights individual pairs to account for overcounting due to some pairs sharing the same source dataset. The causal agent is gpt-3.5-turbo with system message set as “You are a helpful assistant for causal reasoning.”. LMPrior uses davinci-instruct-beta.

# Causal Reasoning and LLM



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

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 diameter.

B. changing the diameter 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: Step 1: Determine which variable is the cause and which is the effect.

In Option A, "age" is the cause and "diameter" is the effect.

In Option B, "diameter" is the cause and "age" is the effect.

Step 2: Consider which variable is more likely to cause a change in the other.

Abalone growth is primarily determined by environmental factors such as water temperature, food availability, and water quality. It is unlikely that changing the age of an abalone would cause its diameter to change significantly. On the other hand, changes in the environment could potentially impact the size of the abalone, and therefore, its diameter could change. Therefore, Option B is more likely: changing the diameter of an abalone causes a change in its age.

<Answer>B</Answer>

(b) Incorrect 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.

# Causal Reasoning and LLM



## 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.
A bird lands in a forest.	What would have happened if a plane had landed in the forest?	The plane would have crashed.;Everything would have been fine.;The plane would have landed safe and sound.;In a forest you will find lots of planes.
A plant grows in a planter.	What would have happened if the planter grows in the plant?	That is not possible.;It would have grown quicker.;The plant would have suffered.;The planter would have cultivated the plant.
A mortician prepares a corpse.	What would have happened if the mortician had prepared a dinner?	He would have had a delicious dish.;Morticians cannot prepare dinners.;The dinner would have been buried.;The mortician would have killed the corpse.
An oil tanker sails across an ocean.	What would have happened if the oil tanker had broken up in an ocean?	There would have been environmental pollution. That is not possible.;The oil tanker would have continued to carry oil.;The oil tanker would have been saved.;
A car crashes into a tree.	What would have happened if the car had parked beneath the tree?	Nothing special would have happened.;The car would have been hit by the tree.;That is not possible.;I think it would have crashed into the tree.
A child draws a picture.	What would have happened if the child had erased the picture?	The picture would not have been visible.;The picture would have been visible.;That is not possible.
A craftsman builds a house.	What would have happened if the house had built a craftsman?	That is not possible.;The house would have been built faster.;Everything would have been fine.;The craftsman would have hands.
A doctor washes their hands at work.	What would have happened if the doctor hadn't washed their hands?	The patients could get an infection.;The patients could get better.;That is not possible.

Table 8: Example scenarios from the CRASS counterfactual reasoning benchmark. The task is to select the best answer choice for the counterfactual question, given a premise.

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	<b>92.44</b>
Human annotators	98.18

# Causal Reasoning and LLM



Model	Accuracy
ada	40.1
text-ada-001	50.0
babbage	50.0
text-babbage-001	50.9
curie	50.0
text-curie-001	50.0
davinci	38.4
text-davinci-001	50.0
text-davinci-002	51.7
text-davinci-003	55.1
gpt-3.5-turbo	71.1
gpt-3.5-turbo (neuropathic pain expert)	75.1
gpt-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)	<b>96.2</b>

Table 4: Accuracy of different versions of GPT on the inferring the edge directions of the Neuropathic pain diagnosis graph. As with the Tübingen dataset, LLMs like gpt-3.5-turbo obtain more than 85% accuracy on determining the direction of edges. The causal agent is gpt-3.5-turbo with a system message set as “You are a helpful assistant for causal reasoning.”



# Causal Reasoning and LLM

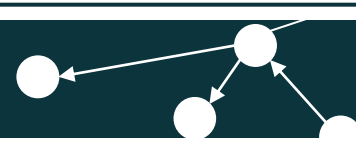


## Inferring necessary or sufficient cause

Vignette Type	Necessary	Sufficient
<i><b>gpt-3.5-turbo</b></i>		
Overdetermination	✓, ✓	X, ✓
Switch	X, X	✓, X
Late preemption	X	X
Early preemption	X, ✓, X	X, X, ✓
Double preemption	✓	✓
Bogus preemption	✓	X,
Short circuit	X	X
Miscellaneous	X, ✓, ✓, X	✓, ✓, X, ✓
Total Accuracy	46.6%	46.6%
<i><b>gpt-4</b></i>		
Overdetermination	✓, ✓	✓, ✓
Switch	✓, ✓	✓, ✓
Late preemption	✓	✓
Early preemption	✓, ✓, ✓	✓,
Double preemption	✓	X
Bogus preemption	✓	✓
Short circuit	X	X
Miscellaneous	✓, X, ✓, ✓	✓, ✓, ✓, ✓
Total Accuracy	<b>86.6%</b>	<b>86.6%</b>

Table 11: Accuracy of gpt-3.5-turbo and gpt-4 on inferring necessary or sufficient cause on 15 standard vignettes. The vignettes are divided into eight types (e.g., Early Preemption type has three vignettes). Each (✓/X) corresponds to a correct/incorrect answer on a single vignette. gpt-3.5-turbo fails at the task (worse than random chance) but gpt-4 can infer necessary and sufficient cause with high accuracy.

# Causal Reasoning and LLM



## Testing Dataset Memorization

Vignette Type	Input Context	Event	Actor	Nec.	Suff.
Overdetermination	Alice (AF) and Bob (BF) each fire a bullet at a window, simultaneously striking the window, shattering it (WS).	window shattering	Alice	No	Yes
Switch	Alice pushes Bob. Therefore, Bob is hit by a truck. Bob dies. Otherwise, Bob would have been hit by a bus, which would have killed him as well.	Bob's death	Alice	No	Yes
Late preemption	Alice (AF) and Bob (BF) each fire a bullet at a window. Alice's bullet hits the window first (AH). The window shatters (WS). Bob's bullet arrives second and does not hit the window (BH).	window shattering	Alice	No	Yes
Early preemption	Suppose Alice reaches out and catches a passing cricket ball. The next thing on the ball's trajectory was a solid brick wall that would have stopped the ball. Beyond that there was a window.	window being intact	Alice	No	Yes
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 intends to put lethal poison into Carol's water. However, Alice does not put lethal poison into Carol's water ( $\neg$ 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 consume the lethal water she would die (CD). Carol consumes her water (CC). Carol does not die ( $\neg$ CD).	Carol's survival	Alice	No	Yes

# Causal Reasoning and LLM



Vignette Type	Necessary	Sufficient
<b><i>gpt-3.5-turbo</i></b>		
Overdetermination	✓, ✓	X, ✓
Switch	X, ✓	✓, X
Late preemption	X	✓
Early preemption	✓, X	X, X
Double preemption	✓	✓
Bogus preemption	✓	X
Short circuit	X	X
Miscellaneous	✓, ✓, ✓, X	✓, X, X, ✓
Total Accuracy	64.2%	42.8%
<b><i>gpt-4</i></b>		
Overdetermination	✓, ✓	✓, ✓
Switch	✓, ✓	X, ✓
Late preemption	✓	✓
Early preemption	✓, ✓	X
Double preemption	✓	✓
Bogus preemption	✓	✓
Short circuit	✓	✓
Miscellaneous	✓, X, ✓, ✓	✓, ✓, X, ✓
Total Accuracy	<b>92.8%</b>	<b>78.5%</b>

Table 12: Testing dataset memorization issues with a novel “lab-vignettes” dataset. The average accuracy of gpt-4 stays the same as in the std vignettes, indicating that gpt-4’s capabilities to infer necessary and sufficient cause can generalize to new data. Inferring necessary cause (93%) emerges as an easier task than inferring sufficient cause (78%).

# Let's Hear Your Thoughts



Do LLMs *Actually* Understand?

# Causal Parrots

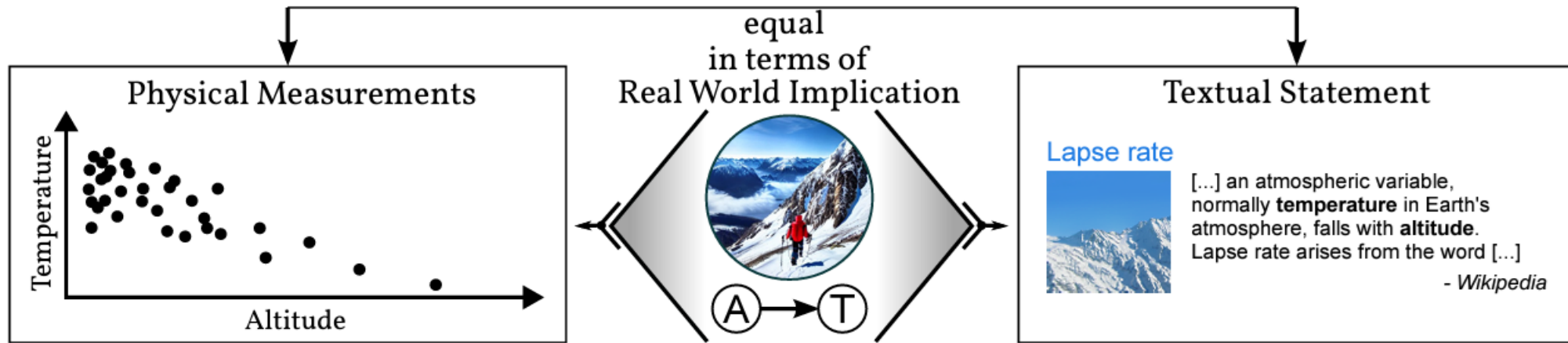


Figure 1: Same Implication, Different Representations.

# Causal Parrots



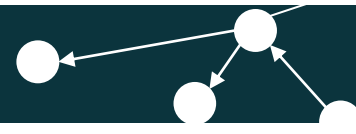
## Formalizing “Correlations of Causal Facts”


- **Causal hierarchy theorem (CHT):** purely observational data collected from a system can not be used to uniquely determine causal statements, when no other causal assumptions are available (Bareinboim et al. 2023)



- No matter how much we scale our foundation models (in terms of data and parameters), we will never be able to perform causal inference
- Causal parrots argue that foundation models might be exploiting a “loop hole” in the CHT
- What happens if the causal assumptions (which are required, by the CHT, for causal inference) are represented in observational data itself?

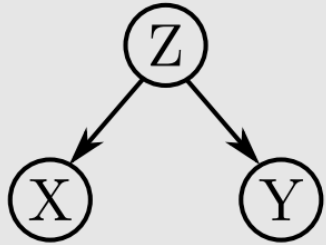
# Causal Parrots



- **Insight 1:** Let  $M$  be some SCM. Knowledge about the structural equations and the causal graph of  $M$  is knowledge about answering  $L_2$  and  $L_3$  queries in  $M$  respectively
  - **Insight 2:** The ‘variables’ of SCMs are not restricted to ‘natural’ concepts, they can be ‘meta’ concepts involving causal facts, that is, knowledge about  $L_2$  and  $L_3$
  - **Meta SCM:** Let  $M_1$  and  $M_2$  be two SCMs such that the observational distribution of  $M_2$  denoted  $L_1(M_2)$  can answer queries w.r.t. the interventional distributions of  $M_1$  denoted  $L_2(M_1)$ , then  $M_2$  is called meta (w.r.t.  $M_1$ )
  - **Correlation of Causal Facts (CCF):** Let  $M_1$  be some SCM and  $M_2$  a respective meta SCM. Further let  $Q \subset L_2(M_1)$  and  $A \subset L_1(M_2)$  be causal queries with their respective answers and  $f$  denotes the LLM’s predictive model. Then we have:  $f(Q) = A \iff f(Q)$  minimizes training error
- 
- In all cases where LLM provides the right causal answer to a causal query, then it is only because (a) this fact was observed in the training data and (b) the correlation with the query is optimal from the perspective of the training objective.

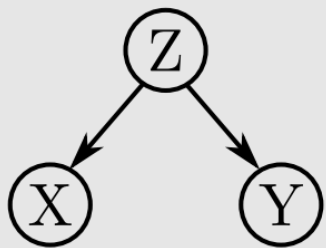


## Causal Assumptions



“Z is common cause of X and Y”  
“X and Y are causally unrelated”

## Causal Assumptions

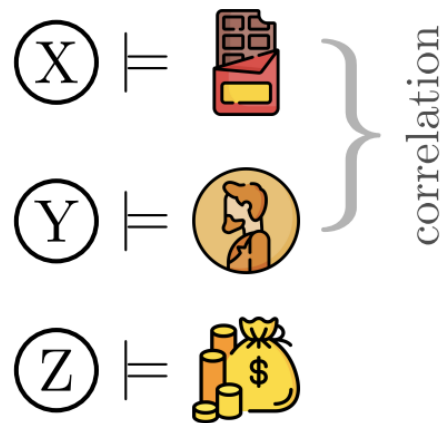





“Z is common cause of X and Y”  
“X and Y are causally unrelated”

## Classical Setting

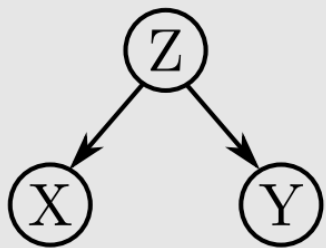
Variables model natural concepts

Example:



Legend:  “Chocolate Consumption”  “Number of Nobel Laureates”  “Gross Domestic Product (GDP)”

## Causal Assumptions

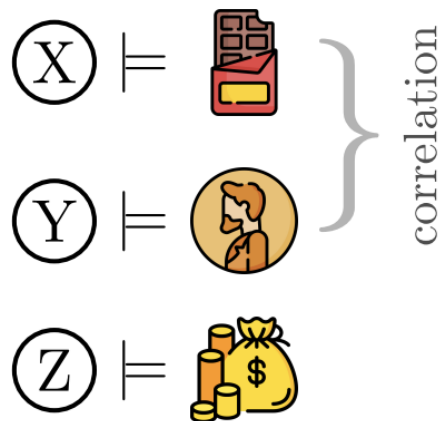


“Z is common cause of X and Y”  
“X and Y are causally unrelated”

## Classical Setting

Variables model natural concepts

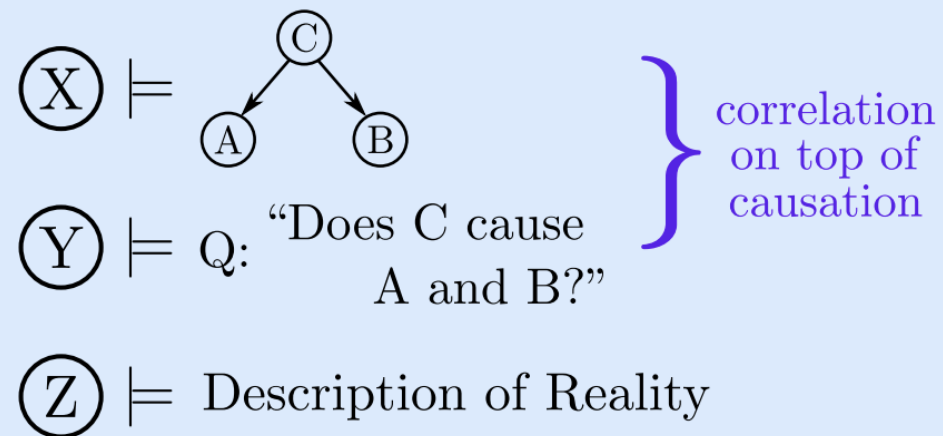
Example:






## Meta-level Setting

Variables model causal assumptions

Example:



Legend:  “Chocolate Consumption”  “Number of Nobel Laureates”  “Gross Domestic Product (GDP)”

# Causal Parrots



*How do LLMs perform..*

*..in “common sense” settings like reasoning or intuitive physics?*

*..in settings where the causal graph is (partially) known?*

*..when using their embeddings of knowledge base facts?*

- Evaluate 3 publicly accessible LLMs
  - OpenAI’s GPT-3
  - AlephAlpha’s Luminous
  - Meta’s OPT

# Causal Parrots



## Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

## Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

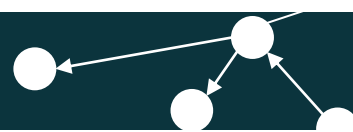
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✅

Regular Prompting vs CoT (Wei et al.)

# Causal Parrots



## “Common Sense” Inference Tasks

	Intuitive Physics						Accuracy
	Rolling (8)	Support (8)	Collisions (4)	Seesaw (4)	Weights (5)	Tools (7)	
GPT-3	<b>6</b>	<b>5</b>	<b>4</b>	<b>2</b>	<b>2</b>	3	<b>61.11%</b>
Luminous	1	0	0	1	1	2	11.11%
OPT	2	0	1	0	0	<b>4</b>	19.44%
GPT-4	<b>7</b>	<b>8</b>	<b>4</b>	<b>3</b>	<b>5</b>	<b>5</b>	<b>91.66% (!)</b>

	Causal Chains (Basic Prop. Logic)									Subchains (4)	Randomized (7)	Accuracy
	N=2	3	4	5	6	7	8	9	10			
GPT-3		✓	✓	✓			✓		✓	2	2	45.00%
Luminous	✓				✓	✓	✓	✓		1	4	50.00%
OPT		✓			✓					0	2	20.00%
GPT-3 (CoT 4,6)	✓	✓	✓	✓	✓	✓	✓	✓	✓	4	<b>7</b>	<b>100.00%</b>
Luminous (CoT 1)	✓	✓	✓	✓	✓	✓	✓	✓	✓	3	3	75.00% *
OPT (CoT 4)	✓	✓	✓	✓	✓	✓	✓	✓	✓	3	4	80.00% *
GPT-4	✓	✓	✓	✓	✓	✓	✓	✓	✓	4	<b>7</b>	<b>100.00% (!)</b>

	Natural Word Chain			Accuracy
	Real World (5)	Imaginary (6)	Mixed (4)	
GPT-4	4	<b>6</b>	3	86.66%
GPT-3	3	0	2	33.33%
Luminous	2	3	2	46.66%
OPT	2	0	2	26.66%
GPT-4 (CoT 3,4)	<b>5</b>	<b>6</b>	<b>4</b>	<b>100.00%</b>
GPT-3 (CoT 2)	<b>5</b>	3	3	73.33%
Luminous (CoT 4)	2	5	2	60.00%
OPT (CoT 1,4)	3, 1	5,6	2,3	66.66%

# Causal Parrots



## Causal Discovery on Ground Truth

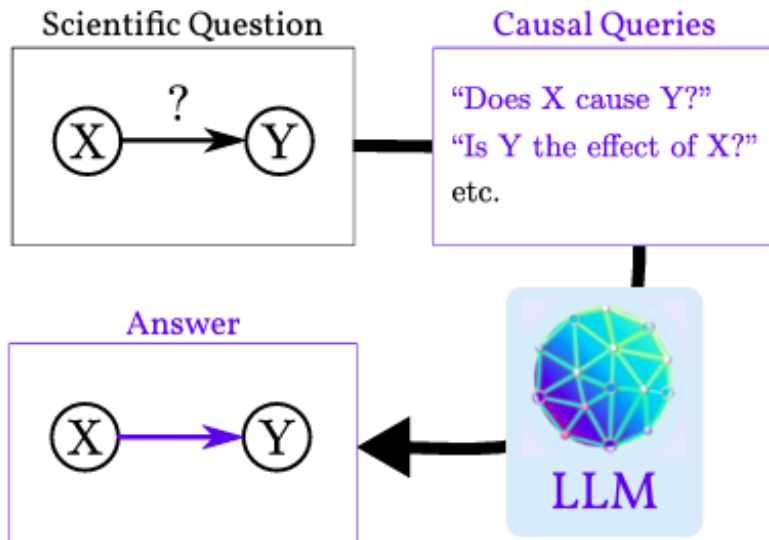


Figure 2: Naïve Causal Discovery with LLMs.

- Query the LLM with all possible combinations of edges between any 2 variables
- 10 questions for Altitude, 100 for Cancer, 60 for Health, 30 for Driving, 100 for Earthquake and 30 for Recovery

# Causal Parrots



## “Common Sense” Inference Tasks

	Metric	Altitude	Health	Driving	Recovery	Cancer	Earthquake	LLM
Causal Graph	SID ↓	<b>0.80</b> $\pm$ 0.40	<b>7.20</b> $\pm$ 0.75	3.00 $\pm$ 0.89	<b>4.00</b> $\pm$ 1.79	11.80 $\pm$ 4.66	<b>11.40</b> $\pm$ 1.50	GPT-3
		1.40 $\pm$ 0.80	9.80 $\pm$ 2.99	<b>2.40</b> $\pm$ 1.20	<b>4.00</b> $\pm$ 2.53	13.20 $\pm$ 7.55	-	GPT-4
		1.20 $\pm$ 0.98	10.60 $\pm$ 1.85	6.00 $\pm$ 0.00	5.40 $\pm$ 1.20	<b>11.40</b> $\pm$ 3.07	16.00 $\pm$ 3.63	Luminous
		1.60 $\pm$ 0.80	10.80 $\pm$ 2.40	5.00 $\pm$ 1.26	5.80 $\pm$ 0.40	16.80 $\pm$ 1.94	15.60 $\pm$ 5.95	OPT
	SHD ↓	0.80 $\pm$ 0.40	<b>4.00</b> $\pm$ 0.63	2.60 $\pm$ 0.49	<b>2.20</b> $\pm$ 0.40	<b>7.00</b> $\pm$ 1.41	<b>4.60</b> $\pm$ 0.80	GPT-3
		0.80 $\pm$ 0.40	6.20 $\pm$ 2.23	<b>1.60</b> $\pm$ 0.80	2.80 $\pm$ 1.60	7.40 $\pm$ 1.62	-	GPT-4
		<b>0.60</b> $\pm$ 0.49	7.00 $\pm$ 1.10	4.20 $\pm$ 0.40	3.40 $\pm$ 0.80	10.00 $\pm$ 3.52	5.60 $\pm$ 1.62	Luminous
		0.80 $\pm$ 0.40	7.40 $\pm$ 1.20	3.40 $\pm$ 1.20	4.00 $\pm$ 0.00	13.20 $\pm$ 1.60	8.60 $\pm$ 3.01	OPT
ML	$F_1$ Score ↑	0.20 $\pm$ 0.40	0.47 $\pm$ 0.14	0.11 $\pm$ 0.23	0.27 $\pm$ 0.33	0.35 $\pm$ 0.11	<b>0.12</b> $\pm$ 0.15	GPT-3
		0.60 $\pm$ 0.33	<b>0.55</b> $\pm$ 0.06	<b>0.64</b> $\pm$ 0.10	<b>0.63</b> $\pm$ 0.19	<b>0.51</b> $\pm$ 0.04	-	GPT-4
		<b>0.80</b> $\pm$ 0.16	0.41 $\pm$ 0.21	0.46 $\pm$ 0.09	0.55 $\pm$ 0.07	0.40 $\pm$ 0.13	0.40 $\pm$ 0.04	Luminous
		0.73 $\pm$ 0.13	0.52 $\pm$ 0.05	0.53 $\pm$ 0.15	0.47 $\pm$ 0.07	0.35 $\pm$ 0.03	0.47 $\pm$ 0.07	OPT
Edges	Sparsity	0.90 $\pm$ 0.20	0.63 $\pm$ 0.28	0.77 $\pm$ 0.31	0.70 $\pm$ 0.31	0.65 $\pm$ 0.16	0.93 $\pm$ 0.07	GPT-3
		0.30 $\pm$ 0.40	0.22 $\pm$ 0.31	0.60 $\pm$ 0.25	0.20 $\pm$ 0.27	0.45 $\pm$ 0.11	-	GPT-4
		0.20 $\pm$ 0.24	0.22 $\pm$ 0.35	0.03 $\pm$ 0.07	0.10 $\pm$ 0.13	0.40 $\pm$ 0.16	0.74 $\pm$ 0.12	Luminous
		0.10 $\pm$ 0.20	0.05 $\pm$ 0.10	0.17 $\pm$ 0.21	0.07 $\pm$ 0.13	0.18 $\pm$ 0.12	0.41 $\pm$ 0.18	OPT
	ADS ↑	0.50	<b>0.62</b>	<b>0.33</b>	0.50	0.69	0.00	GPT-3
		0.50	0.61	0.17	<b>0.83</b>	<b>0.85</b>	-	GPT-4
		<b>1.00</b>	0.53	0.17	0.17	0.38	0.26	Luminous
		0.50	0.25	0.25	0.33	0.28	<b>0.47</b>	OPT

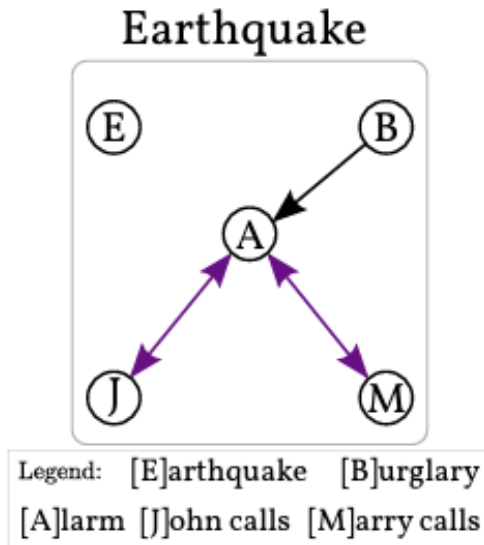
Table 2: Comparing LLMs prediction to existing ground truth causal structures. The metrics concerned with the causal graph structure (SID, SHD) reveal a closer match of GPT-3 and GPT-4 predictions to the ground truth causal structures than for the other LLMs. High  $F_1$  Scores and low sparsity indicate densely connected graph prediction by Luminous and OPT. This can be desired for ML applications. The ADS reveals that all LLMs increase their decisiveness on edge directions when querying with asymmetric sentence templates. Metrics for GPT-4 on Earthquake are not computed to prevent skewed results due to unclear judgement of meta answers.



# Causal Parrots



## Meta Answers



- GPT-4 gives meta-answers for the Earthquake data set
- Variables 'John calls' and 'Marry calls' refer to specific persons for which no further information is available
- Either no connection is predicted or a meta-answer (purple) is given for all queries involving 'John' or 'Marry'

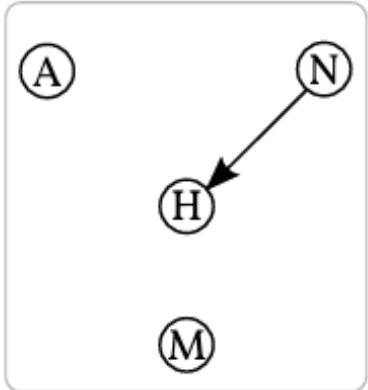
Figure 3: Meta answers for unknown concepts.

# Causal Parrots

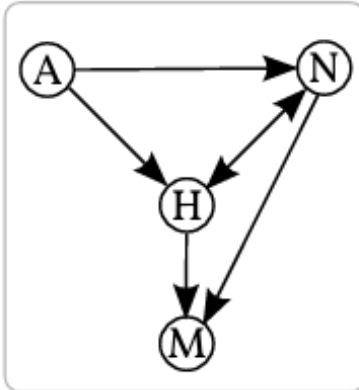


## Sensitivity

“Does  $X$  cause  $Y$ ?”



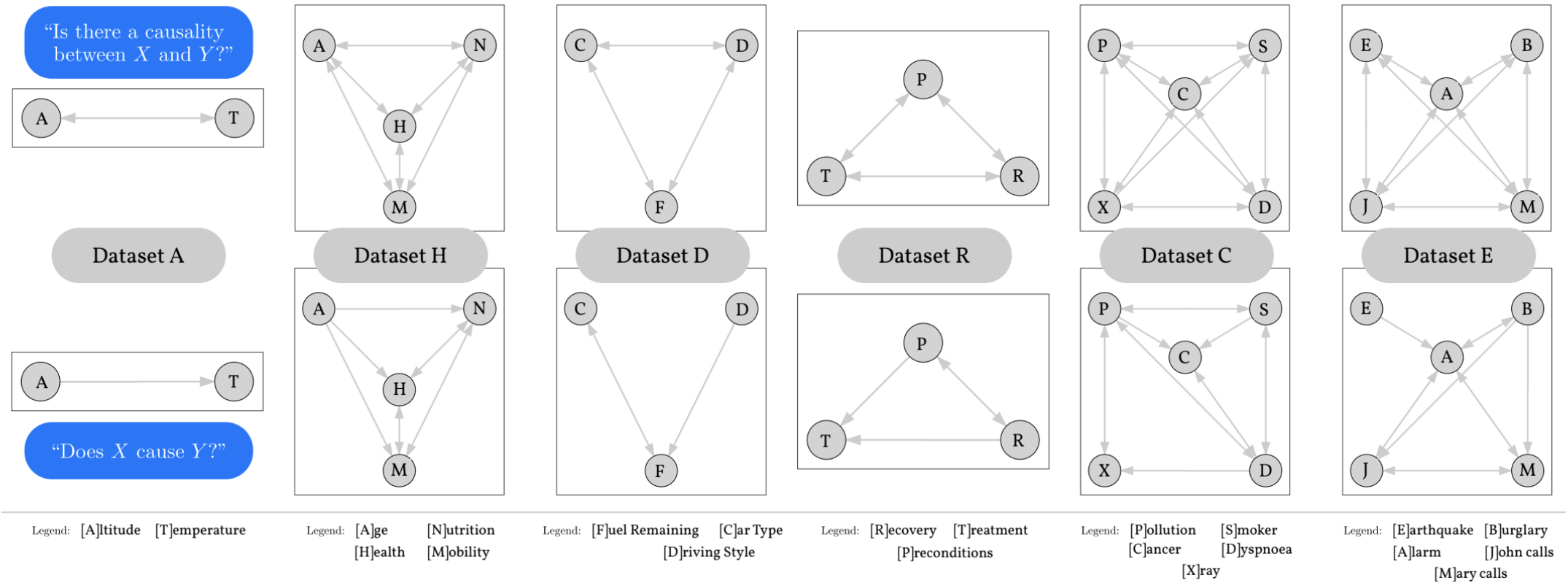
“Is there a causality between  $X$  and  $Y$ ?”



Legend: [A]ge [N]utrition [H]ealth [M]obility

- A possible interpretation is that a keyword such as ‘causality’ might be embedded further away from an alternate keyword (here for instance ‘cause’) within the LLM’s latent space, thus answering correctly
- the results we present do not contradict prior literature but rather complement further understanding of LLM capabilities
- observing a high accuracy in the Tübingen cause-effect pairs data set

Figure 4: Sensitivity to Query Wording.



**Figure 3: Asymmetric Query Wording Implies Unidirectedness.** Language FM naive graph predictions on data sets that provide a causal graph (FM-O is shown). Top row, predictions with a symmetric query wording, bottom row, predictions with an asymmetric query wording. Surprisingly, the FM is capable of deciding multiple edges uniquely (and correctly) when switching to the asymmetric formulation without explicit guarantees to such behavior.

So, Are LLMs Causal?

Let's Ask ChatGPT

Large language models (LLMs) like GPT-4 are not inherently causal. To understand why, it's important to distinguish between correlation and causation, as well as to explore the nature of LLMs:

### 1. Correlation vs. Causation:

- **Correlation:** When two variables are related, but one does not necessarily cause the other. For example, ice cream sales and drowning incidents may be correlated because they both increase in the summer, but buying ice cream does not cause drowning.
- **Causation:** When one variable directly affects another. For example, smoking causes an increase in the risk of lung cancer.

## 2. Nature of Large Language Models:

- **Training on Data:** LLMs are trained on large datasets using statistical methods to learn patterns, correlations, and structures in the data. They do not understand or infer causality from the data itself.
- **Prediction and Generation:** These models generate text based on the learned patterns and statistical relationships. They predict the next word or sequence of words that are most likely given the input context.
- **Lack of Explicit Causal Reasoning:** LLMs do not have an explicit understanding of causal mechanisms. They do not model the underlying processes that cause one event to lead to another. Their responses are based on the probability distributions derived from the training data.

### 3. Causal Models vs. LLMs:

- **Causal Models:** These involve explicit representations of cause-and-effect relationships, often using tools like causal diagrams (e.g., Bayesian networks) and methods from causal inference (e.g., do-calculus).
- **LLMs:** These are based on patterns and correlations in data without explicit causal reasoning.



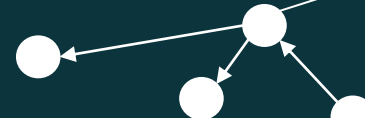
## Current Efforts and Future Directions:

- Researchers are exploring ways to incorporate causal reasoning into LLMs. This includes integrating causal models or using methods from causal inference to enhance the reasoning capabilities of LLMs.
- **Hybrid Models:** Combining LLMs with causal models could potentially lead to systems that understand and reason about causality while leveraging the rich linguistic capabilities of LLMs.

In summary, while large language models excel at capturing and utilizing correlations in data to generate coherent and contextually relevant text, they do not inherently possess causal reasoning capabilities. They are correlation-based models, not causal ones.

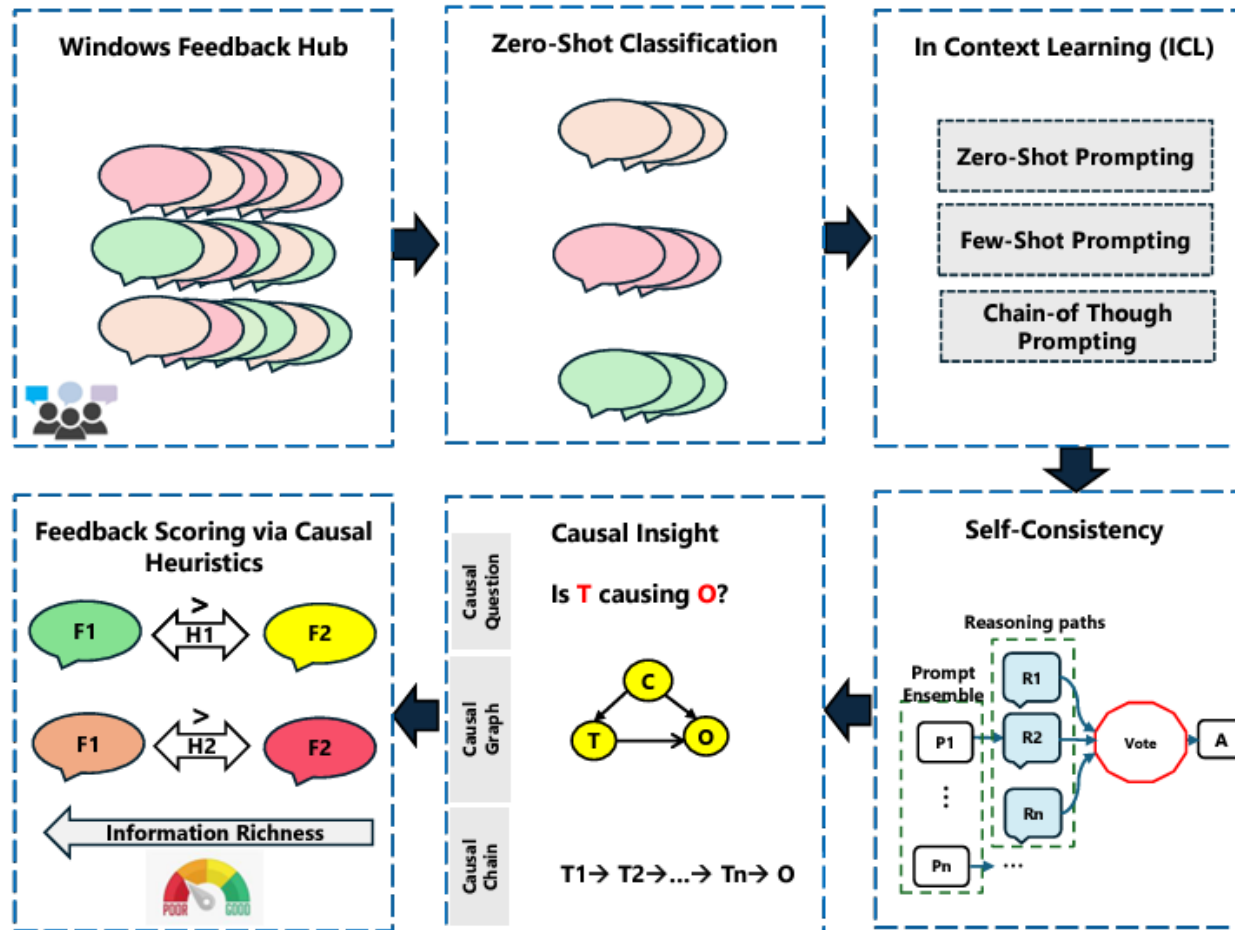
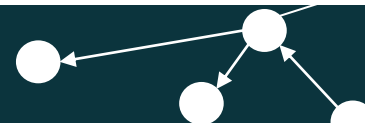
# Applications

# Self-Consistent Causal Insights

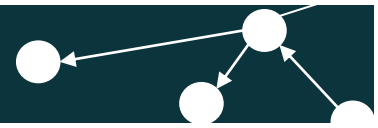


- Leverage Double Machine Learning (DML) to associate users' feedback on Microsoft Windows Feedback Hub with telemetry signals
  - training causal models which are able to create unbiased estimates of causal effects
  - far less prone to overfitting
- Necessity of domain knowledge for model design
- Large Language Models (LLMs) are used to generate a prior model
  - could be used as a heuristic for measuring feedback informativeness
- In-Context Learning (ICL) to design a modified self-consistency framework to mitigate LLMs' hallucination to extract reliable causal variables
- extract chains of events from feedback to provide engineers with a "causal summary" of reported bugs

# Self-Consistent Causal Insights



# Self-Consistent Causal Insights



## Zero-Shot Chain-of-Thought

Extract treatments, outcomes and confounders that are being discussed in the following text, which is delimited by triple backticks.  
treatments are variables that cause some outcomes.  
Outcomes are effects of the treatments.  
Confounders are variables that affect both outcome and treatment.  
Create a list of treatment and effect tuples.  
Format your response as a list of tuples with first element as treatment and second as outcome and third as confounders in the form of (treatment, outcome,[confounders])  
**let's think step by step.**  
**Text sample:** ""(feedback)""

## Reasoning path 1

- **Treatment:** Modern Standby mode
- **Outcome:** TextInputHost.exe crashes
- **Confounders:** Microsoft Windows Operating System, unknown factors that may impact stability while in sleep mode
- [(Modern Standby mode, TextInputHost.exe crashes, [Microsoft Windows Operating System, unknown factors])]

## Reasoning path 2

- **Treatments:** - Modern Standby mode
  - **Outcomes:** - TextInputHost.exe constantly crashes during sleep mode -
  - **Confounders:** - None mentioned in the text
- List of treatment and effect tuples:
- (Modern Standby mode, TextInputHost.exe constantly crashes during sleep mode, [])

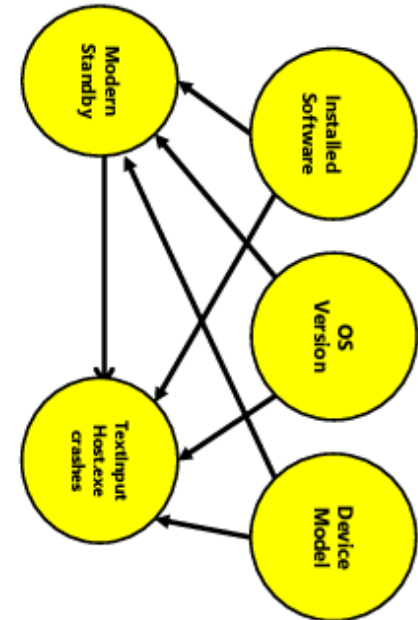
## Reasoning path 3

- List of treatment and effect tuples:
- (Modern Standby, constant crashing of TextInputHost.exe, [unknown potential confounders])
  - (TextInputHost.exe, laptop crashing on sleep mode, [potential confounders include laptop model, operating system version, other software installed])

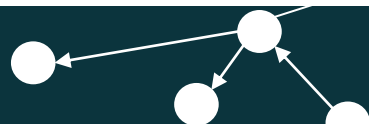
## Reasoning path 4

- List of treatments, outcomes and confounders:
- **Treatment:** Modern Standby sleep mode
  - **Outcome:** TextInputHost.exe crashes
  - **Confounders:** Microsoft Windows Operating System critical events
- List of tuples:
- (Modern Standby sleep mode, TextInputHost.exe crashes, [Microsoft Windows Operating System critical events])

## Causal Graph



# Self-Consistent Causal Insights



## Few-Shot Chain-of-Thought

Extract treatments, outcomes and confounders that are being discussed in the following text, which is delimited by triple backticks.

treatments are variables that cause some outcomes.

Outcomes are effects of the treatments. Confounders are variables that affect both outcome and treatment.

Create a list of treatment and effect tuples.

Format your response as a list of tuples with first element as treatment and second as outcome and third as confounders in the form of (treatment, outcome, [confounders])

**let's think step by step.**

**Example:**

**Feedback:** on windows 11, my laptop, While in sleep mode, drains a lot of battery and the fans turn on randomly.

**Answer:** List of treatments, outcomes and confounders:

- Treatment: sleep mode
- Outcome: draining a lot of battery, fan turns on randomly
- Confounders: Laptop model, OS build,

List of tuples:  
- (sleep mode, draining a lot of battery, fan turns on randomly, [Laptop model, OS build])

**Feedback:** ""(feedback)""

**Answer:**

## Reasoning path 1

List of treatments, outcomes and confounders:

- **Treatment:** Modern Standby sleep mode
- **Outcome:** constant crashing of TextInputHost.exe
- **Confounders:** Microsoft Windows Operating System, specific laptop model, other currently running applications

**List of tuples:**

- (Modern Standby sleep mode, constant crashing of TextInputHost.exe, [Microsoft Windows Operating System, specific laptop model, other currently running applications])

## Reasoning path 2

List of treatments, outcomes and confounders:

- **Treatment:** Modern Standby (sleep mode)
- **Outcome:** TextInputHost.exe crashes
- **Confounders:** Laptop model, OS build, other applications running in the background

**List of tuples:**

- (Modern Standby, TextInputHost.exe crashes, [Laptop model, OS build, other applications running in the background])

## Reasoning path 3

List of treatments, outcomes and confounders:

- **Treatment:** Modern Standby (sleep mode)
- **Outcome:** TextInputHost.exe crashes
- **Confounders:** Laptop model, OS build, other applications running in the background

**List of tuples:**

- (Modern Standby, TextInputHost.exe crashes, [Laptop model, OS build, other applications running in the background])

## Reasoning path 4

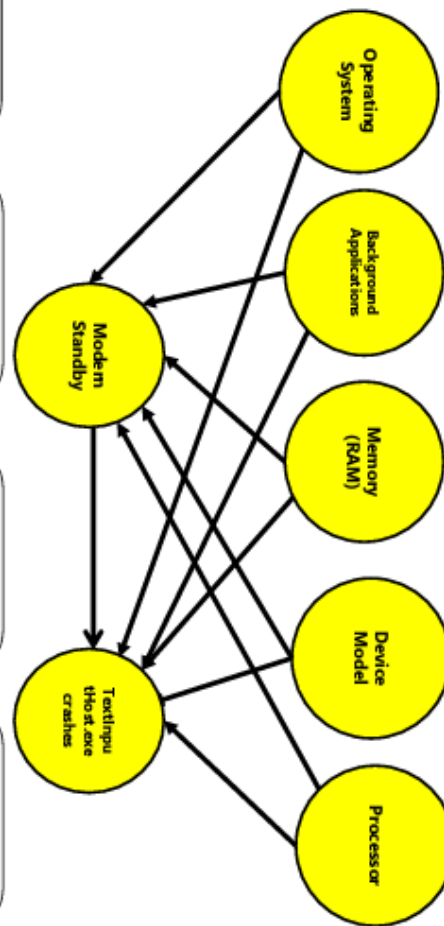
List of treatments, outcomes and confounders:

- **Treatment:** modern standby (sleep mode)
- **Outcome:** TextInputHost.exe constantly crashes
- **Confounders:** Operating System, technical specifications such as RAM and processor

**List of tuples:**

- (modern standby (sleep mode), TextInputHost.exe constantly crashes, [Operating System, RAM, processor])

## Causal Graph



# Causal Language in Science Claims

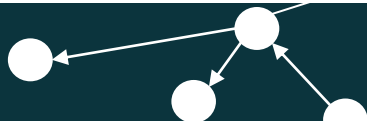


- Evaluate ChatGPT's ability to understand causal language in science papers and news
- Testing its accuracy in a task of labeling the strength of a claim:
  - Causal
  - Conditional Causal
  - Correlational
  - No Relationship
- ChatGPT is still behind the existing fine-tuned BERT models by a large margin
- Difficulty understanding conditional causal claims
- Chain-of-thought prompting was faithful and helpful for improving prompt performance
- Finding the optimal prompt is difficult with inconsistent results

# Commonsense Reasoning



# ChatGPT: A Good Causal Reasoner?



## Event Causality Identification (ECI)

**Input:** Minutes after a woman was suspended and escorted from her job at the Kraft Foods plant in Northeast Philadelphia, she returned with a gun and opened fire, killing two women and critically injuring a third co-worker before being taken into custody.

**Question:** is there a causal relationship between "suspended" and "injuring" ?

**Answer:** Yes

## Causal Discovery (CD)

*Multiple Choice*

**Input Event:** The man fell unconscious.

**Question:** Please select the cause of the input event from the following options.

**Option 1:** The assailant struck the man in the head.

**Option 2:** The assailant took the man's wallet.

**Answer:** Option 1

*Binary Classification*

**Event A:** The man fell unconscious.

**Event B:** The assailant struck the man in the head.

**Question:** is there a causal relationship between Event A and Event B ?

**Answer:** Yes

## Causal Explanation Generation (CEG)

**Cause:** The assailant struck the man in the head.

**Effect:** The man fell unconscious.

**Question:** why the cause can lead to the effect?

**Answer:** Hit to head caused brain disruption, leading to unconsciousness.

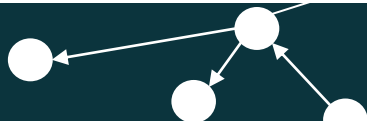
# ChatGPT: A Good Causal Reasoner?



## Used Datasets

- EventStoryLine v0.9 (ESC)
  - 22 topics, 258 documents, 5,334 events and 1,770 causal event pairs
- Causal-TimeBank (CTB)
  - 184 documents, 6,813 events and 318 causal event pairs
- COPA
  - classic dataset for causal reasoning
  - 1,000 MCQs that primarily focus on everyday life scenarios
- e-CARE
  - 21,324 multiple-choice questions covering a wide range of domains

# ChatGPT: A Good Causal Reasoner?



Methods	ESC			CTB			MAVEN-ERE		
	P	R	F1	P	R	F1	P	R	F1
<b>BERT-Base (Devlin et al., 2019)</b>	38.1	56.8	45.6	41.4	45.8	43.5	52.5	75.6	61.9
<b>RoBERTa-Base (Liu et al., 2019)</b>	42.1	64.0	50.8	39.9	60.9	48.2	<b>52.8</b>	75.1	<b>62.0</b>
<b>KEPT (Liu et al., 2023)</b>	50.0	68.8	57.9	48.2	60.0	53.5	-	-	-
<b>DPJL (Shen et al., 2022)</b>	<b>65.3</b>	70.8	<b>67.9</b>	<b>63.6</b>	66.7	<b>64.6</b>	-	-	-
<b>text-davinci-002</b>	23.2	80.0	36.0	5.0	75.2	9.3	19.6	<b>92.9</b>	32.4
<b>text-davinci-003</b>	33.2	74.4	45.9	8.5	64.4	15.0	25.0	75.1	37.5
<b>gpt-3.5-turbo</b>	27.6	80.2	41.0	6.9	82.6	12.8	19.9	85.8	32.3
<b>gpt-4</b>	27.2	<b>94.7</b>	42.2	6.1	<b>97.4</b>	11.5	22.5	92.4	36.2
	Pos	Neg	Full	Pos	Neg	Full	Pos	Neg	Full
<b>BERT-Base (Devlin et al., 2019)</b>	59.5	<b>83.6</b>	79.7	46.4	<b>87.8</b>	<b>86.2</b>	75.7	86.9	85.1
<b>RoBERTa-Base (Liu et al., 2019)</b>	63.8	82.8	<b>79.8</b>	62.3	86.4	85.5	76.6	<b>87.1</b>	<b>85.4</b>
<b>text-davinci-002</b>	80.0	43.1	49.6	75.2	41.9	43.2	<b>92.9</b>	21.2	33.5
<b>text-davinci-003</b>	74.4	67.7	68.9	64.4	71.9	71.6	75.1	53.6	57.2
<b>gpt-3.5-turbo</b>	80.2	54.4	59.0	82.6	55.0	56.0	85.8	28.5	38.3
<b>gpt-4</b>	<b>94.7</b>	41.4	51.4	<b>97.4</b>	39.1	41.4	92.4	33.9	44.0

Table 1: Experimental results (%) on the ECI task. P, R and F1 indicate Precision, Recall and F1-score, respectively. Pos, Neg and Full indicate accuracy on the causal pairs, non-causal pairs and all test datas, respectively.

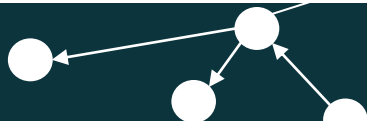
# ChatGPT: A Good Causal Reasoner?



Methods	Multiple Choice		Binary Classification					
	e-CARE	COPA	e-CARE			COPA		
	Full	Full	Pos	Neg	Full	Pos	Neg	Full
<b>BERT-Base (Devlin et al., 2019)</b>	75.4	75.4	65.7	95.2	65.0	59.5	88.9	54.5
<b>RoBERTa-Base (Liu et al., 2019)</b>	70.7	80.5	64.6	74.6	64.5	66.0	69.0	63.5
<b>text-davinci-002</b>	78.4	94.4	18.5	<b>95.2</b>	56.8	55.6	<b>92.4</b>	74.0
<b>text-davinci-003</b>	76.7	93.2	41.0	86.4	63.7	80.4	82.3	<b>81.4</b>
<b>gpt-3.5-turbo</b>	79.1	96.3	75.5	66.9	71.2	96.3	43.2	69.7
<b>gpt-4</b>	<b>84.5</b>	<b>98.1</b>	<b>84.8</b>	57.5	<b>71.2</b>	<b>97.9</b>	38.5	68.2

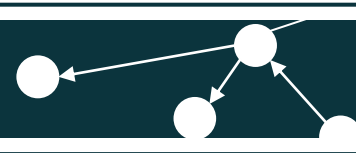
Table 2: Experimental results (%) on the CD task. Pos, Neg and Full indicate accuracy on the causal pairs, non-causal pairs and all test datas, respectively.

# ChatGPT: A Good Causal Reasoner?



Methods	e-CARE		
	AVG-BLEU	ROUGE-1	Human Evaluation
<b>GRU-Seq2Seq (Chung et al., 2014)</b>	18.7	21.3	0.0
<b>GPT2 (Radford et al., 2019)</b>	32.0	31.5	20.0
<b>LLaMA 7B (Touvron et al., 2023)</b>	<b>40.3</b>	37.1	63.0
<b>FLAN-T5 11B (Chung et al., 2022)</b>	38.5	<b>42.8</b>	66.0
<b>text-davinci-003</b>	10.55	37.49	83.0
<b>gpt-3.5-turbo</b>	7.32	40.31	82.0
<b>gpt-4</b>	6.47	39.77	85.0
<b>Human Generation (Du et al., 2022)</b>	35.51	33.46	<b>89.5</b>

Table 3: Experimental results (%) on the CEG task. “Human Generation” is the human annotated explanations provided by Du et al. (2022).



## Commonsense Causality Reasoning

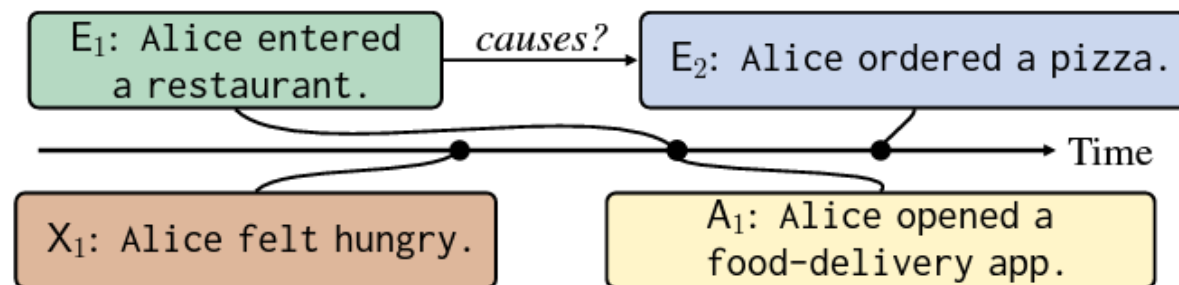


Figure 1: **An Example of CCR:** *does  $E_1$  cause  $E_2$ ?* The temporal order  $E_1 \prec E_2$  does not necessitate causation due to confounding co-occurrences (e.g.,  $X_1$ ). Since when *conditioning on  $X_1$* , a *comparable* intervention  $A_1$  of  $E_1$  also precedes  $E_2$ , the effect from  $E_1$  to  $E_2$  shrinks.

reasoning about cause-and-effect relationships between events in natural language descriptions



- Precedence does not imply causation, which warns us post-hoc fallacies
- Causation implies precedence
- The events must be compared with those that are in pari materia or having balanced covariates
- CCR formulation in terms of temporality has several benefits
  - the intrinsic temporality of causal principles characterizes its central role in CCR
  - temporal signals bring about incidental supervision
  - reasoning temporality has witnessed decent progress lately, making it more accessible than directly detecting causal signals

# ROCK

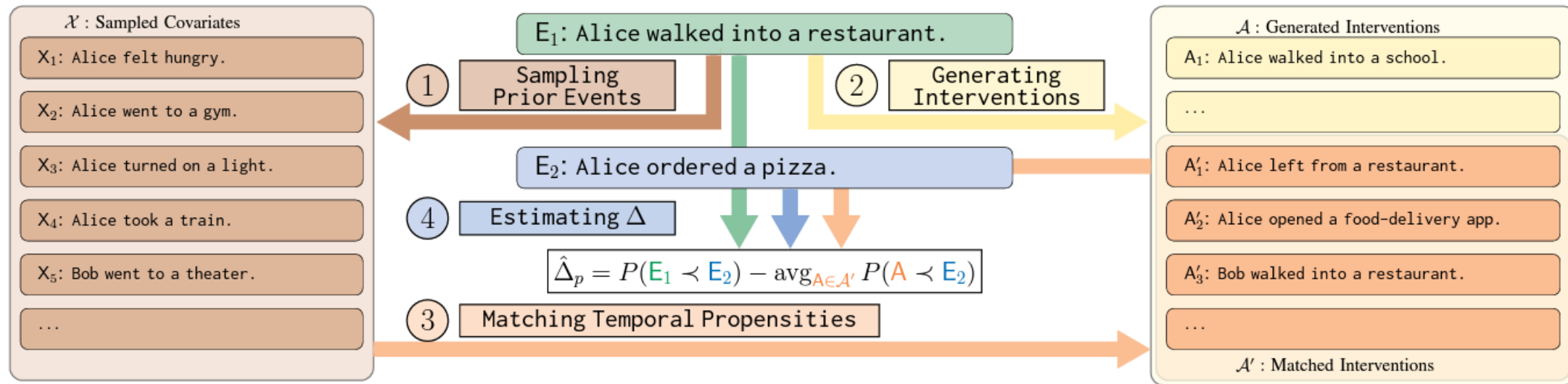
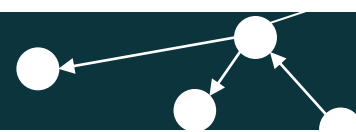


Figure 2: **Illustration of the ROCK framework.** Does  $E_1$  cause  $E_2$ ? To answer this query, ① the event sampler samples a set of covariates  $\mathcal{X}$  of events  $X_k$  that occur preceding  $E_1$ . ② The intervention generator generates a set  $\mathcal{A}$  of interventions  $A_k$  on  $E_1$ . ③ A subset  $\mathcal{A}' \subset \mathcal{A}$  of interventions is selected whose temporal propensities  $q(\mathbf{x}; A)$  is close to that of  $E_1$ ,  $q(\mathbf{x}; E_1)$  (Equation (7)). ④ The temporal predictor uses  $\mathcal{A}'$  to estimate  $\Delta$ .

$$\left\{ \begin{aligned} \hat{\Delta}_p &= f(E_1, E_2) - \frac{1}{|\mathcal{A}'|} \sum_{A \in \mathcal{A}'} f(A, E_2), \\ \mathcal{A}' &:= \left\{ A \in \mathcal{A} : \frac{1}{|\mathcal{X}|} \|q(\mathbf{x}; A) - q(\mathbf{x}; E_1)\|_p \leq \epsilon \right\}. \end{aligned} \right.$$





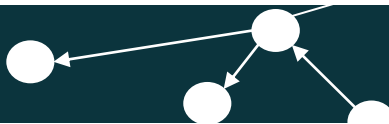
- Intervention generator generates a set of events that are considered as interventions of the event
- Achieved by masking these components individually and filling in the masks using an LM.

	Random Baseline	$\hat{\Delta}_1 \uparrow$ $L_1$ -Balanced	$\hat{\Delta}_2 \uparrow$ $L_2$ -Balanced	$\hat{\Delta}_{E_1} \uparrow$ Temporal	$\hat{\Delta}_{\mathcal{A}} \uparrow$ Unbalanced	$\hat{\Delta}_{\mathcal{X}} \uparrow$ Misspecified
COPA-DEV	$0.5 \pm 0.050$	0.6900	<b>0.7000</b>	0.5800	0.5600	<b>0.5300</b>
COPA-TEST	$0.5 \pm 0.022$	<b>0.5640</b>	<b>0.5640</b>	<b>0.5200</b>	0.5400	0.5240
GLUCOSE-D1	$0.5 \pm 0.040$	0.6645	<b>0.6968</b>	0.5677	<b>0.5742</b>	0.6581
COPA-DEV (-T)	$0.5 \pm 0.050$	0.6200	<b>0.6300</b>	0.5300	<b>0.4800</b>	0.5300
COPA-TEST (-T)	$0.5 \pm 0.022$	<b>0.5800</b>	0.5740	<b>0.4540</b>	0.4600	0.4860
GLUCOSE-D1 (-T)	$0.5 \pm 0.040$	0.6065	<b>0.6194</b>	0.5548	0.4387	<b>0.3742</b>

Table 1: **Best zero-shot results.** Shaded rows have temporal fine-tuning (T) disabled. (i) Estimators with temporal propensities balanced ( $\hat{\Delta}_1$  and  $\hat{\Delta}_2$ ) perform consistently better than the unbalanced and the temporal estimators. (ii) In general, without temporality fine-tuning (“-T”, see Section 4), the performances degrade.

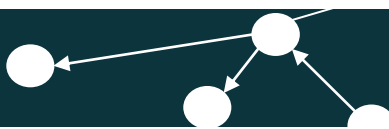
# Fairness and Bias

# Word Embeddings via Causal Inference

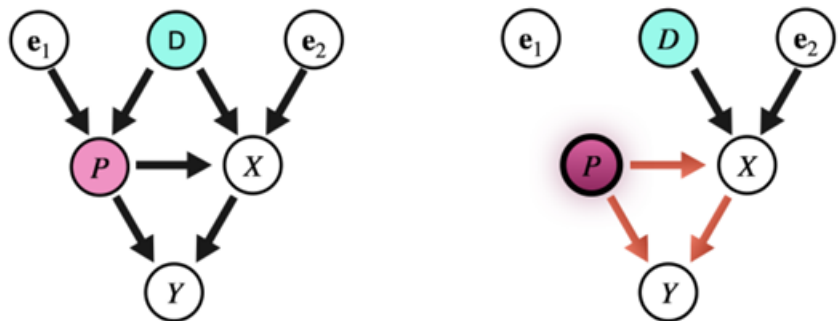


- Word embeddings trained on human-generated corpora have strong gender biases
- Previous work mainly on modeling bias and only implicitly consider semantic information
- Overlooks the complex underlying causal structure among bias and semantic components
- Model intervention techniques to determine causal effects in a causal model
- Apply the model intervention to proxy variables of the gender bias rather than the gender bias variables

# Word Embeddings via Causal Inference



**Definition 1** (Potential proxy bias.) A variable  $Y$  in a causal graph exhibits potential proxy bias if there exists a directed path from  $D$  to  $Y$  that is blocked by a proxy variable  $P$  and if  $Y$  itself is not a proxy.

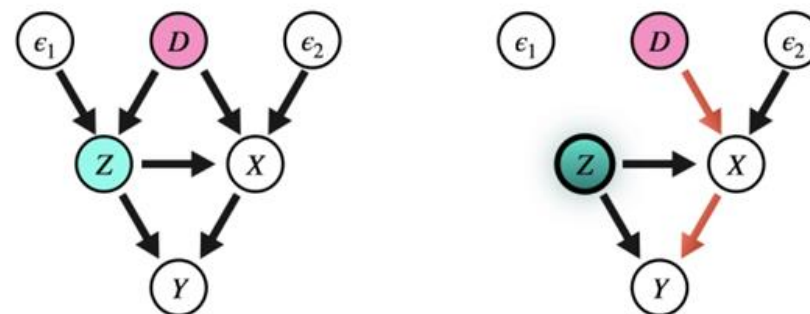


(a) Proxy bias

(b) Intervention on proxy bias

Figure 1: A causal graph for proxy bias removal.

**Definition 2** (Unresolved bias.) A variable  $Y$  in a causal graph exhibits unresolved bias if there exists a directed path from  $D$  to  $Y$  that is not blocked by a resolving variable  $Z$  and  $Y$  itself is non-resolving.

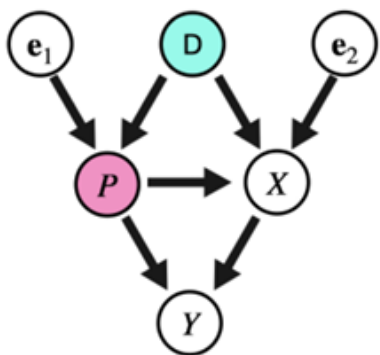
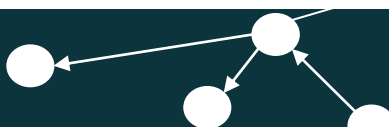


(a) Unresolved bias

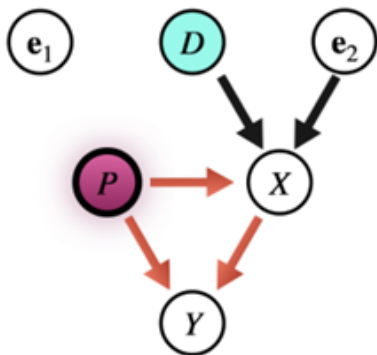
(b) Intervention on unresolved bias

Figure 2: A causal graph for unresolved bias removal.

# Word Embeddings via Causal Inference



(a) Proxy bias



(b) Intervention on proxy bias

Figure 1: A causal graph for proxy bias removal.

---

Algorithm 1: (P-DeSIP) Removing potential proxy bias.

---

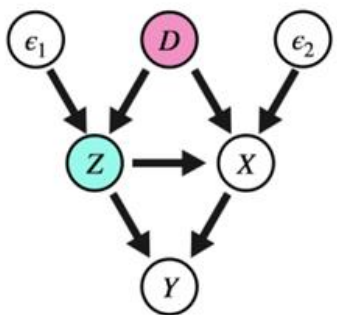
**Input:**  $\mathbf{D}$ ,  $\mathbf{P}$ ,  $\mathbf{X}$ , and  $\mathbf{Y}$ .

- 1: Solve  $\mathbf{X} = \mathbf{D}\alpha_1 + \mathbf{P}\alpha_2 + e_2$  by PLS to get  $(\hat{\alpha}_1, \hat{\alpha}_2)$
- 2: Solve  $\mathbf{Y} = \mathbf{P}\beta_1 + \mathbf{X}\beta_2$  by PLS to get  $(\hat{\beta}_1, \hat{\beta}_2)$
- 3: Compute  $\hat{\mathbf{Y}} = (\mathbf{X} - \mathbf{P}\hat{\alpha}_2)\hat{\beta}_2$
- 4: Compute  $\hat{\mathbf{Y}}_{\Delta^\perp} = \mathbf{Y} - \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{Y}$
- 5: Compute  $\hat{\mathbf{Y}}_{\text{P-DeSIP}} = \hat{\mathbf{Y}} + \hat{\mathbf{Y}}_{\Delta^\perp}$

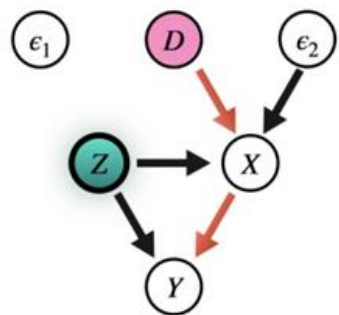
**Output:**  $\hat{\mathbf{Y}}_{\text{P-DeSIP}}$  as debiased word matrix.

---

# Word Embeddings via Causal Inference



(a) Unresolved bias



(b) Intervention on unresolved bias

---

Algorithm 2: (U-DeSIP) Removing unresolved bias.

---

**Input:**  $D$ ,  $Z$ ,  $X$ , and  $Y$ .

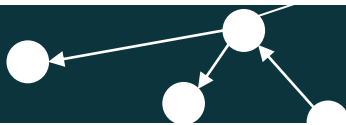
- 1: Solve  $\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta}_1 + \mathbf{X}\boldsymbol{\theta}_2$  by PLS to get  $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$
- 2: Compute  $\hat{\mathbf{Y}} = \mathbf{Z}\hat{\boldsymbol{\theta}}_1$
- 3: Compute  $\hat{\mathbf{Y}}_{\Delta^\perp} = \mathbf{Y} - \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{Y}$
- 4: Compute  $\hat{\mathbf{Y}}_{\text{U-DeSIP}} = \hat{\mathbf{Y}} + \hat{\mathbf{Y}}_{\Delta^\perp}$

**Output:**  $\hat{\mathbf{Y}}_{\text{U-DeSIP}}$  as debiased word matrix.

---

Figure 2: A causal graph for unresolved bias removal.

# Causal View of Entity Bias in LLMs



**Context:** Bill Gates went to Microsoft Building 99.

**Question:** What's the relation between Bill Gates and Microsoft in the given context?

**Option:** founder, visitor.

**Answer** with one word: founder (GPT-3.5) ❌

*Assume subject\_entity can be any of Bill Gates, Jeff Bezos, and Steve Jobs, while object\_entity can be any of Google, Microsoft, and Meta.*

**Context:** subject\_entity went to object\_entity Building 99.

**Question:** What's the relation between subject\_entity and object\_entity in the given context?

**Option:** founder, visitor.

**Answer** with one word: visitor (GPT-3.5) ✅

Entity bias occurs when models overly rely on prediction shortcuts triggered by specific entities to make spurious predictions

Figure 1: An example of entity bias in GPT-3.5. Our in-context intervention mitigates the conflicts between parametric knowledge and contextual knowledge.

# Causal View of Entity Bias in LLMs

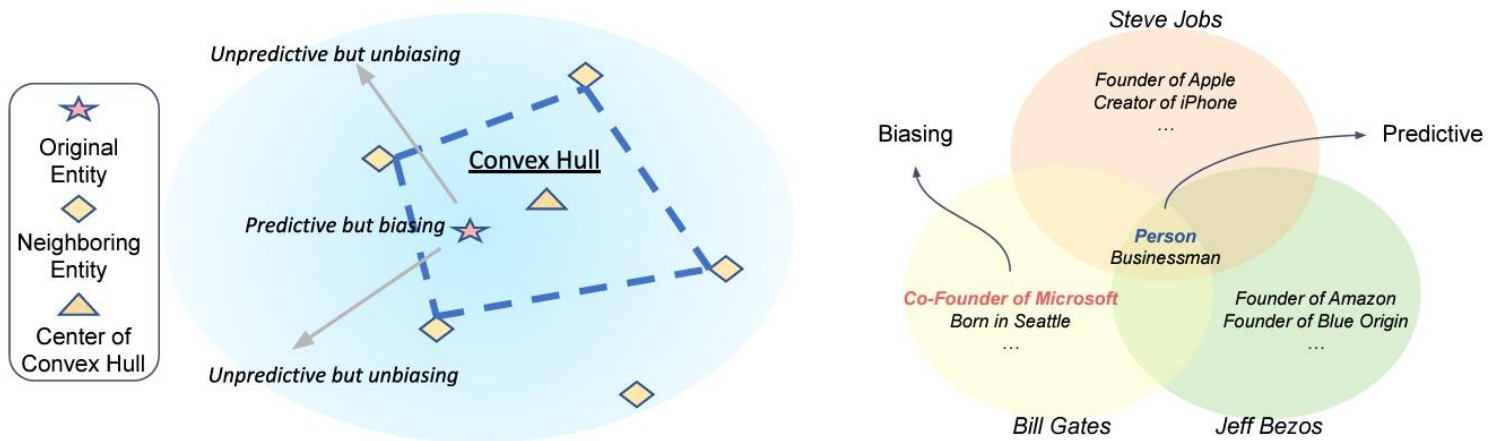
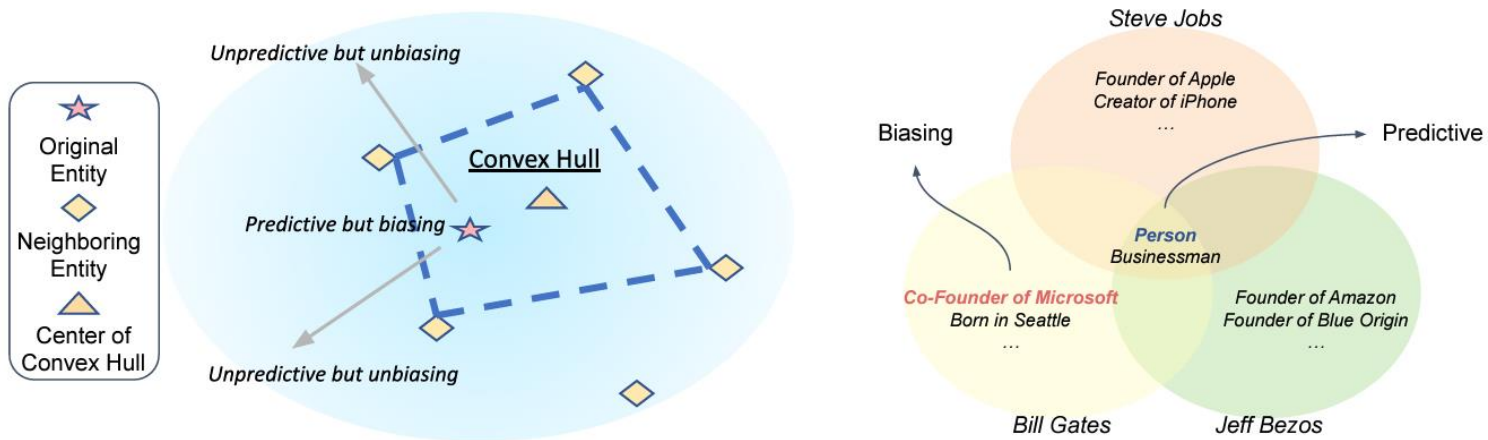
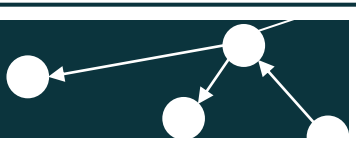


Figure 3: Left: Training-time intervention with  $k = 4$ . Right: Example of predictive and biasing information.

- For each entity presented in the input text, first find its top  $k$  nearest neighbors
- Then construct the smallest convex hull to cover the original entity and neighboring entities
- The entity specific biasing information, which has the potential to trigger spurious shortcuts, gradually diminishes from the original entity towards the border of the convex hull



# Causal View of Entity Bias in LLMs



- For white-box models of which the parameters are accessible, we can effectively address their internal bias through training-time intervention
- Realize the causal intervention by perturbing the input entities or entity tokens using their neighboring counterparts in the embedding space

Figure 3: Left: Training-time intervention with  $k = 4$ . Right: Example of predictive and biasing information.

# Causal View of Entity Bias in LLMs



## 1. Replace entities with placeholders

*Context*

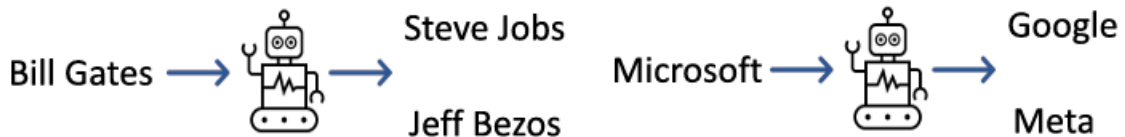
**subject\_entity** went to **object\_entity** Building 99.

*Question*

What's the relation between **subject\_entity** and **object\_entity** in the given context?

## 2. Find similar entities

Name two persons (organizations) similar to Bill Gates (Microsoft)



## 3. Define placeholders with examples

*Definition*

Assume that **subject\_entity** can be any of Bill Gates, Jeff Bezos, and Steve Jobs, while **object\_entity** can be any of Google, Microsoft, and Meta.

## 4. Query with the new prompt

*Definition*

*Context*

*Question*

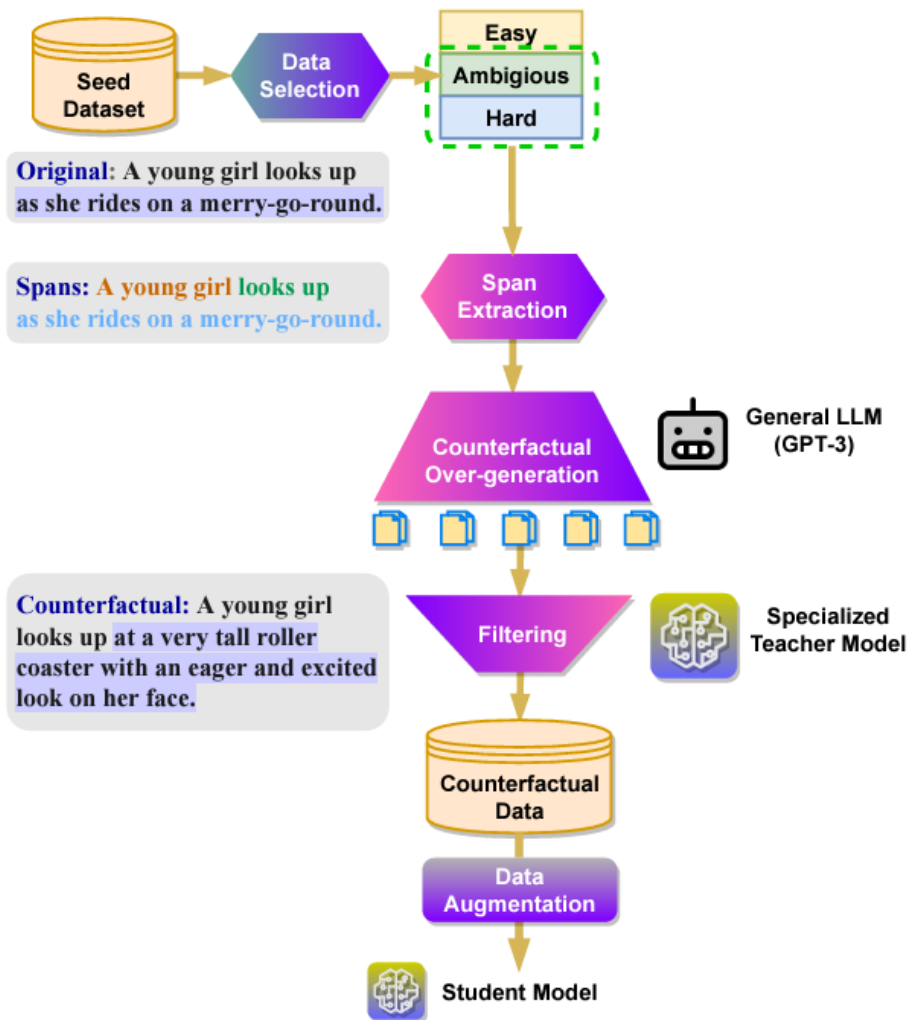
*Options*



visitor

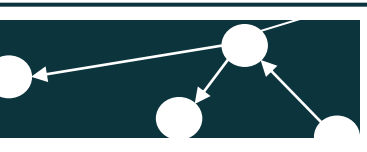
# LLMs for Causal Inference

# DIStilled COUNTERfactual Data (DISCO)

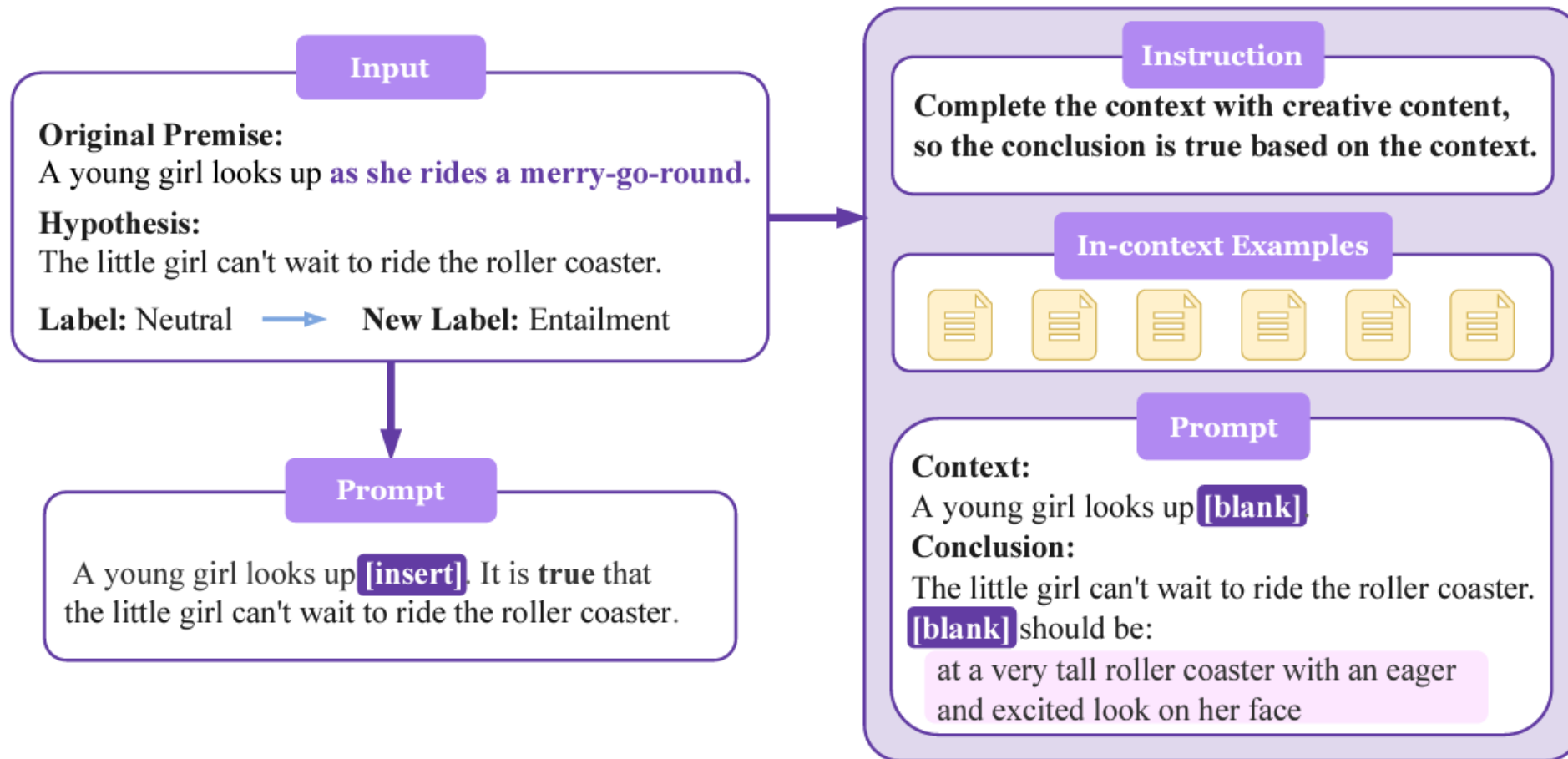


**Central Idea:** prompt a large language model to generate perturbations that can flip the current label to a new one

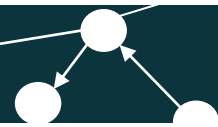
# DIStilled COunterfactual Data (DISCO)



## Perturbation Generation Process



# DIStilled COunterfactual Data (DISCO)



Method	Size	Model Robustness							OOD Generalization		
		PI-CD	PI-SP	IS-CS	LI-LI	LI-TS	ST	HANS	MNLI <sup>1</sup>	MNLI <sup>2</sup>	QNLI
<b>Large-size augmentation on full SNLI</b>											
SNLI	549,367	82.2	69.0	68.4	93.6	72.5	72.4	73.1	78.5	78.2	64.5
+ WANLI	652, 252	83.4	<u>82.7</u>	69.5	86.2	84.3	67.4	<u>87.4</u>	78.2	78.0	<u>78.6</u>
+ Z-aug	<b>1,142,475</b>	<b>84.1</b>	72.5	<u>72.6</u>	<b>93.9</b>	<u>87.1</u>	<u>75.4</u>	68.3	<u>80.0</u>	<b>80.7</b>	75.0
<b>Augmentation on subset of SNLI</b>											
SNLI-subset	100,000	82.0	71.7	65.1	85.5	83.9	69.5	65.8	78.0	79.1	73.4
+ Tailor	192,457	79.5	52.0	55.8	84.6	80.1	62.7	55.8	64.1	65.7	71.4
+ Human-CAD	108,330	82.8	<u>77.8</u>	69.2	90.7	87.1	71.3	65.5	79.0	79.0	72.8
+ <i>DISCO</i> (ours)	165,418	<b>84.1</b>	74.1	<b>73.5</b>	<u>92.1</u>	<u>88.4</u>	<b>77.0</b>	<u>70.1</u>	<b>80.5</b>	<u>80.2</u>	<u>77.7</u>
<b>Augmentation on WANLI</b>											
WANLI	102,885	65.6	81.3	65.9	65.6	82.7	56.5	<b>89.4</b>	76.1	76.3	81.1
+ <i>DISCO</i> (ours)	177,885	<u>82.8</u>	<b>83.8</b>	<u>72.0</u>	<u>86.8</u>	<u>85.1</u>	<u>68.6</u>	87.4	<u>80.0</u>	<u>78.7</u>	<b>81.4</b>
<b>Trained on <i>DISCO</i> (ours) data only</b>											
<i>DISCO</i> (ours)	<b>75,000</b>	83.5	77.4	73.3	89.4	<b>88.9</b>	76.3	70.7	79.2	79.5	79.1

Table 3: Results on Stress-tests, robust NLI test suites (Liu et al., 2020b), MNLI-hard, and QNLI. The bold numbers are the highest accuracy within a column, and the underlined numbers are the highest accuracy for each section. MNLI<sup>1</sup> refers to MNLI-hard-match, and MNLI<sup>2</sup> refers to MNLI-hard-mismatch.

# Causal Inference via LLM Guided Discovery

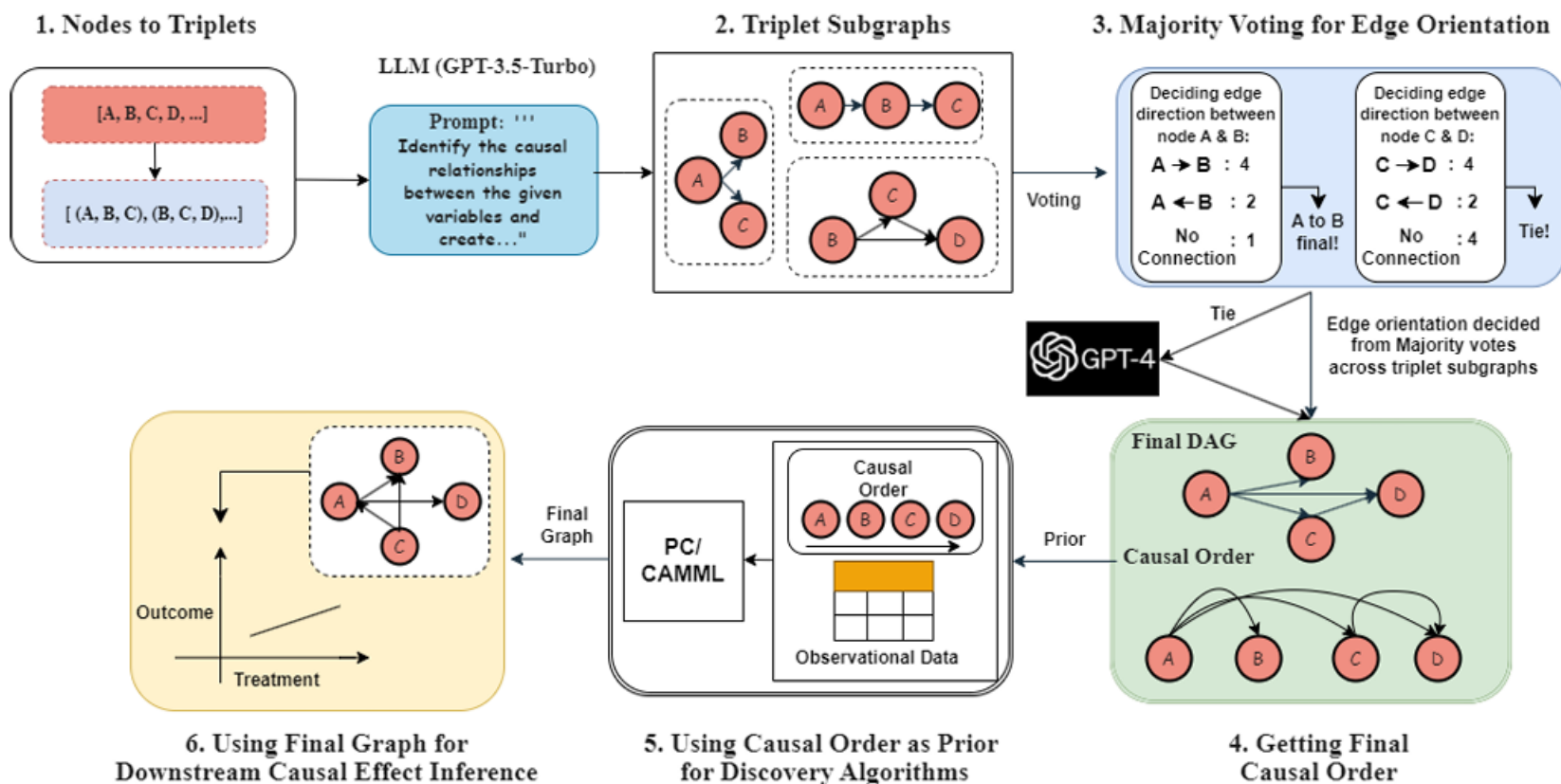


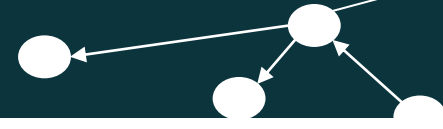
Figure 1: The *LLM-augmented* causal inference process based on inferring causal order. We propose a triplet-based prompting technique to infer all three-variable subgraphs and aggregate them using majority voting to produce a causal order. The causal order can then be used to identify a valid backdoor adjustment set. Ties in causal order are broken using another LLM (e.g., GPT-4). For robustness, LLM-generated causal order may be used in conjunction with discovery algorithms like PC or CaMML.

# Causal Inference via LLM Guided Discovery



	Dataset	PC	SCORE	ICA LINGAM	Direct LiNGAM	NOTEARS	CaMML	Ours (PC+LLM)	Ours (CaMML+LLM)
$N = 250$	Earthquake	0.30±0.44	4.00±0.00	3.20±0.39	3.00±0.00	1.80±0.74	2.00±0.00	<b>0.00±0.00</b>	<b>0.00±0.00</b>
	Cancer	<b>0.00±0.00</b>	3.00±0.00	4.00±0.00	3.60±0.48	2.00±0.00	2.00±0.00	<b>0.00±0.00</b>	<b>0.00±0.00</b>
	Survey	0.50±0.00	3.00±0.00	6.00±0.00	6.00±0.00	3.20±0.39	3.33±0.94	<b>0.00±0.00</b>	3.33±0.94
	Asia	2.33±0.59	5.00±0.00	6.20±0.74	7.00±0.00	4.00±0.00	1.85±0.58	0.00±0.00	<b>0.97±0.62</b>
	Asia-M	2.00±0.00	5.00±0.00	7.60±0.48	6.20±1.16	3.40±0.48	1.00±0.00	<b>0.00±0.00</b>	1.71±0.45
	Child	8.16±1.58	8.80±2.70	12.8±0.97	13.0±0.63	15.0±1.09	<b>3.00±0.00</b>	4.00±0.00	3.53±0.45
	Neuropathic	3.25±0.00	6.00±0.00	13.0±6.16	10.0±0.00	9.00±0.00	10.4±1.95	<b>1.00±0.00</b>	5.00±0.00
$N = 500$	Earthquake	0.85±0.65	4.00±0.00	3.20±0.39	3.40±0.48	1.20±0.40	<b>0.00±0.00</b>	0.4±0.89	<b>0.00±0.00</b>
	Cancer	<b>0.00±0.00</b>	3.00±0.00	3.40±0.48	3.00±0.00	2.00±0.00	1.00±0.00	<b>0.00±0.00</b>	1.00±0.00
	Survey	1.75±0.00	4.00±0.00	6.00±0.0	6.00±0.00	3.40±0.48	3.39±0.08	<b>1.00±0.00</b>	3.33±0.94
	Asia	3.00±0.00	5.00±0.00	5.60±0.48	7.00±0.00	3.20±0.39	3.81±0.39	1.00±0.00	<b>0.97±0.62</b>
	Asia-M	2.00±0.00	6.00±0.00	7.60±0.48	5.00±0.00	3.80±0.39	2.00±0.00	1.00±0.00	<b>0.17±0.45</b>
	Child	9.79±1.17	6.20±1.32	12.2±0.74	10.6±1.35	15.4±0.48	<b>2.00±0.00</b>	4.6±1.34	<b>2.00±0.00</b>
	Neuropathic	7.50±0.00	6.00±0.00	9.00±1.41	13.0±0.00	11.0±0.00	<b>5.32±0.57</b>	8.00±0.00	7.49±0.64
$N = 1000$	Earthquake	0.80±0.57	4.00±0.00	3.00±0.00	3.00±0.00	1.00±0.00	<b>0.00±0.00</b>	0.20±0.44	<b>0.00±0.00</b>
	Cancer	<b>0.00±0.00</b>	3.00±0.00	3.00±0.00	3.00±0.00	2.00±0.00	1.60±0.48	<b>0.00±0.00</b>	<b>0.00±0.00</b>
	Survey	1.00±0.00	4.00±0.00	5.80±0.39	5.40±0.48	3.20±0.39	2.71±0.27	<b>1.00±0.00</b>	2.83±0.00
	Asia	3.09±1.05	4.00±0.00	6.20±0.74	6.60±0.48	3.40±0.48	1.75±0.43	1.75±0.95	<b>0.97±0.62</b>
	Asia-M	2.50±0.00	4.00±0.00	8.00±0.00	5.20±0.39	3.40±0.48	2.04±0.51	2.00±0.00	<b>0.65±0.47</b>
	Child	9.61±1.07	3.80±0.74	12.2±1.72	11.8±0.74	15.2±0.97	<b>2.00±0.00</b>	8.0±0.00	<b>2.00±0.40</b>
	Neuropathic	-	6.00±0.00	<b>4.00±0.81</b>	12.0±0.00	12.0±0.00	5.54±0.75	-	10.1±2.12
$N = 5000$	Earthquake	0.30±0.67	4.00±0.00	2.80±0.39	3.00±0.00	1.00±0.00	0.80±0.97	<b>0.00±0.00</b>	<b>0.00±0.00</b>
	Cancer	<b>0.00±0.00</b>	3.00±0.00	3.00±0.00	3.00±0.00	2.00±0.00	2.00±0.00	<b>0.00±0.00</b>	<b>0.00±0.00</b>
	Survey	2.00±0.00	4.00±0.00	5.00±0.00	5.00±0.00	3.00±0.00	3.33±0.69	<b>2.00±0.00</b>	2.60±0.00
	Asia	3.05±0.94	4.00±0.00	6.60±0.79	4.40±1.35	3.40±0.48	1.75±0.43	2.80±1.30	<b>0.97±0.62</b>
	Asia-M	1.00±0.00	4.00±0.00	7.60±0.48	4.60±0.48	3.20±0.39	1.68±0.46	0.20±0.44	<b>0.00±0.00</b>
	Child	8.42±0.75	3.00±0.00	12.6±0.79	10.8±1.72	14.2±0.40	<b>3.00±0.00</b>	7.00±0.00	<b>3.00±0.00</b>
	Neuropathic	9.00±0.00	6.00±0.00	9.33±0.94	10.0±0.00	10.0±0.00	4.20±0.96	9.00±0.00	<b>1.23±0.42</b>
$N = 10000$	Earthquake	<b>0.00±0.00</b>	4.00±0.00	3.00±0.00	3.00±0.00	1.00±0.00	0.40±0.48	<b>0.00±0.00</b>	<b>0.00±0.00</b>
	Cancer	<b>0.00±0.00</b>	3.00±0.00	3.00±0.00	3.00±0.00	2.00±0.00	0.60±0.80	<b>0.00±0.00</b>	<b>0.00±0.00</b>
	Survey	2.00±0.00	4.00±0.00	5.00±0.00	5.00±0.00	3.00±0.00	3.60±1.35	2.00±0.00	<b>1.83±0.00</b>
	Asia	1.95±0.41	4.00±0.00	6.00±0.00	4.40±1.35	3.00±0.00	1.40±0.48	1.20±0.83	<b>0.34±0.47</b>
	Asia-M	1.75±0.00	4.00±0.00	8.00±0.00	4.80±0.39	3.00±0.00	2.00±0.00	<b>0.00±0.00</b>	<b>0.00±0.00</b>
	Child	7.67±0.65	3.00±0.00	12.2±1.46	11.6±0.48	14.4±0.48	2.80±0.84	7.00±0.00	<b>1.00±0.00</b>
	Neuropathic	10.00±0.00	6.00±0.00	<b>1.00±0.00</b>	10.0±0.00	10.0±0.00	3.00±0.00	10.00±0.00	<b>1.00±0.00</b>





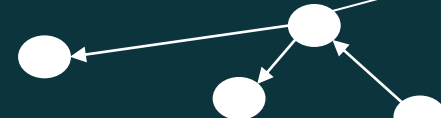
## Definition 4 (Prior Constraints on BN Structures)

*Common forms of constraints on the BN structure are as follows:*

- 1) Edge existence,  $x \rightarrow y$ : There must be a directed edge  $(x, y)$  in the BN;*
- 2) Edge forbidden,  $x \not\rightarrow y$ : There must not be a directed edge  $(x, y)$  in the BN;*
- 3) Order constraint,  $x < y$ : The variable  $x$  occurs before the variable  $y$ ;*
- 4) Ancestral constraint,  $x \rightsquigarrow y$ : There must be a directed path from the variable  $x$  to the variable  $y$ .*

Here authors use LLM-driven statements as the ancestral constraint

# LLMs for Advanced Causal Discovery



## Identifying Causal Statements

Table 1: A practical example of prompts with respect to the three-stage LLM causal discovery framework.

---

### Prompt Understand:

You are an expert on *Domain*. You are investigating the cause-and-effect relationships between the following variables in your filed. Variable abbreviations and values are presented as follows. Please understand the real meaning of each variable according to their possible values, and explain them in order.

variable  $sym_1$ , values  $val_1$

...

variable  $sym_n$ , values  $val_n$

---

### Prompt Causal Discovery:

Based on the meaning of variables you provide, analyze the cause-and-effect relationships between them. Please give the results as a directed graph network in the form of  $\langle edge \rangle variable1 \rightarrow variable2 \langle \backslash edge \rangle$ . Make sure that each edge represent a direct causality between the two variables.

---

### Prompt Revsion:

Based on your explanation, check whether the following causal statements are correct, and give the reasons.

$x_{i_1} \rightarrow x_{j_1}$

...

$x_{i_m} \rightarrow x_{j_m}$

---

help LLMs comprehend the meaning of variables

require the direct nature of the causality, as all indirect causality can be inferred from direct causality

# Final Discussion: Are LLMs Causal?

Thank you for  
staying with us  
throughout this  
course



Any suggestions/comments: [d.s.dhami@tue.nl](mailto:d.s.dhami@tue.nl) or [adele.ribeiro@uni-marburg.de](mailto:adele.ribeiro@uni-marburg.de)