Causal Parrots

Large Language Models May Talk Causality But Are Not Causal

Zečević et al. TMLR 2023

Alycia N. Carey
Department of Electrical Engineering and Computer Science
University of Arkansas

Trustworthy ML Reading Group
February 7, 2024

Slides *generously* adapted from Aneesh Komanduri

Conduct - Carter

- From Ethos of MLC…
  - https://mlcollective.org/wiki/code-of-conduct/

- Highlights
  - Expectation of Confidentiality
  - Reporting -> send me (Carter) a direct message

Motivation

"These models are castles in the air. They have no foundations whatsoever."

-- Judea Pearl

- Foundation models:
  - Lack foundation
  - Unexplainable
  - regurgitation at scale
  - Diminishing returns? Is bigger better?
    - On the dangers of Stochastic Parrots – Bender et al.

https://ichthyoid.writeas.com/castle-in-the-sky-a-study-in-world-building

Question: Can LLMs perform rigorous causal reasoning?

TL;DR: NO

Question: Why do they seem to answer causal questions?

TL;DR: Encounter *correlations* over causal facts during training

CORRELATION ≠ CAUSATION

Causal Reasoning

The process of *identifying* causality: the relationship between a cause and its effect.

- One of our most central cognitive competencies which enables us to adapt to the world
- E.g., predict future events, or diagnose the causes of observed facts
- Without causal reasoning, we would not have made progress in various empirical sciences (physics, medicine, biology,…)
- A central topic of philosophy throughout history

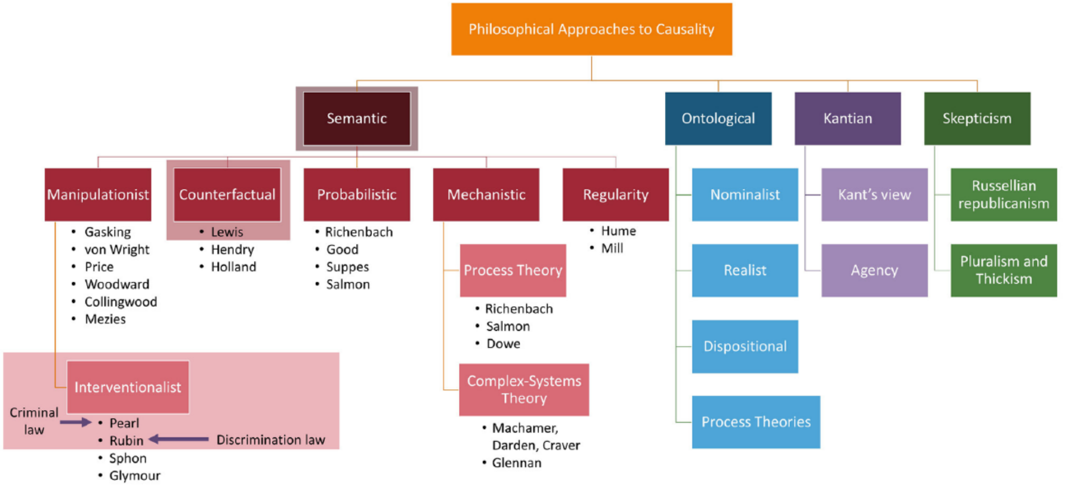


FIGURE 6 | The multiple different approaches to causality. The first level indicates the types of questions that can be asked about causality and the rows below are answers to the questions. For example, probabilistic approaches are an answer to semantic questions. The shadow-boxed items are elements discussed in this publication: interventionalist approaches of Pearl and Rubin as well as the counterfactual groundings of both based on Lewis' work. Additionally, we map where two different types of U.S. legal cases fall within the causal framework. We refer interested readers to Broadbent (2020) for details on the approaches not covered.

Carey and Wu. The Causal Fairness Field Guide: Perspectives from Social and Formal Sciences. Frontiers in Big Data. 2022

Plato’s Allegory of the Cave

https://jamesian58.blogspot.com/2021/02/five-lessons-from-allegory-of-cave.html

“To which extent can you learn about the real world’s functioning by just observing the shadows of it’s objects?”

Pearl's Causal Hierarchy

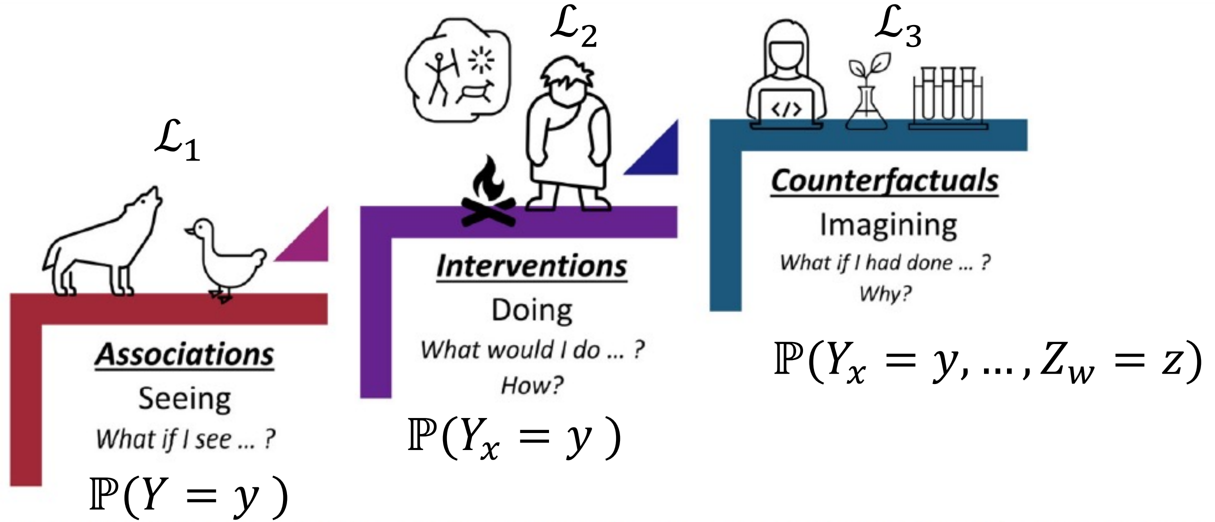


FIGURE 3 | Pearl's Ladder of Causation. The first rung, associations, only allows predictions based on passive observations. The second rung, interventions, not only relies on seeing, but also changing what is. Rung three, counterfactuals, deals with the imaginary, or what might have been.

Carey and Wu. The Causal Fairness Field Guide: Perspectives from Social and Formal Sciences. Frontiers in Big Data. 2022

Example: Altitude vs. Temperature

Fundamental difference between “understanding” a fact and simply “knowing” it

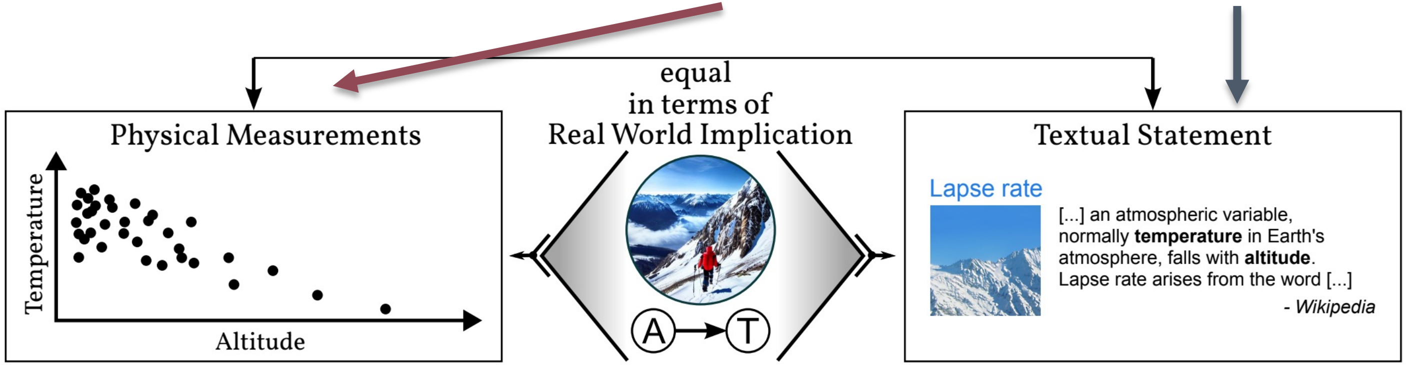


Fig. 1. **Same Implication, Different Representations.** When we consider the causal relationship between altitude (A) and temperature (T), then it is apparent that given the laws of physics we have an increase in altitude leading to a decrease in temperature. Graphically we can depict the relationship as A→T, whereas the actual ‘increase-decrease’ relationship can only be specified through the SCM formalism with its structural equations, that is some f such that $T = f(A, U)$ where U are exogenous variables. The ground truth SCM underlying our laws of physics generates observational data in the form of numerical tuples (a, t) as seen on the left scatter plot. To infer the casual relation, we can resort to algorithms for causal discovery. However, crucially, the same knowledge achieved through such induction can be represented within *text* for ‘free’ as one simply recites the Wikipedia article found on the right. While the article on the right is correct, and thus represents a fact about the actual world, there is no such guarantee for arbitrary other texts. That is, a model that simply obtains its knowledge from various Wikipedia statements will also learn untrue statements, statements that are not facts, thus explaining behavior that is correct sometimes and wrong other times.

Pearl's Causal Hierarchy Theorem

- CHT guarantees that purely observational data collected cannot be used to uniquely determine causal statements (given no other assumptions)
  - Cannot answer L3 questions with L1 data
  - *No matter how much we scale our foundation models we will never be able to perform causal inference*
- LLM pretrained on purely observational data and *seem* to do causal reasoning
- But doesn't this violate Pearl's hierarchy? In theory, yes.
- LLMs might be exploiting a loophole in Pearl's Causal Hierarchy to "talk" causal
- **What if causal assumptions required to answer causal statements are embedded in the *observational distribution*?**
  - E.g., training data

Structural Causal Model (SCM)

A **Structural Causal Model (SCM)** is a triple $\langle \mathbf{Z}, \mathbf{N}, \mathbf{F} \rangle$ such that

- $\mathbf{Z}$ is the set of endogenous (internal) variables,
- $\mathbf{N}$ is the set of exogenous (external) noise variables,
- $\mathbf{F}$ is a collections of structural equations of the form:
  $Z_j = f_j(\mathbf{Pa}_j, N_j), j = 1, \ldots, n$
  - $\mathbf{Pa}_j \subseteq \mathbf{Z} \setminus \{Z_j\}$ are the causal parents of Z_j and the N_j are jointly independent noise variable

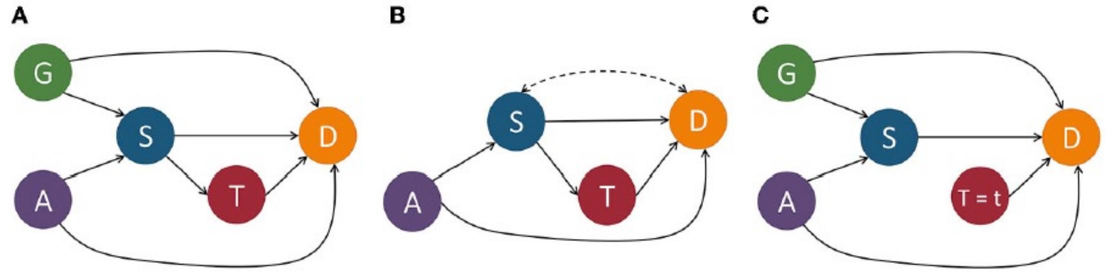


FIGURE 2 | Example of a structural causal model depicting the relationships between variables that determine the survival of some disease. *G*, gender; *A*, age; *S*, severeness; *T*, treatment; *D*, survival. In **(A)**, we include the exogenous variable gender and in **(B)** we show it removed by adding a bi-directional dashed line between *S* and *D* since this is a semi-Markovian model. In **(C)**, we perform an intervention on variable *T* by setting it equal to *t*.

Carey and Wu. The Causal Fairness Field Guide: Perspectives from Social and Formal Sciences. Frontiers in Big Data. 2022

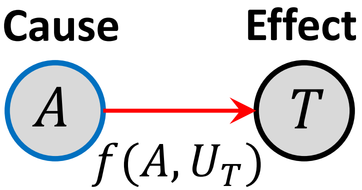
Meta SCM

Can we model SCMs describing physical phenomena using an SCM?

- Variables of SCMs are not restricted to ‘natural’ concepts, they can be ‘meta’ concepts involving causal facts

2-variable SCM $\mathcal{M}_1$

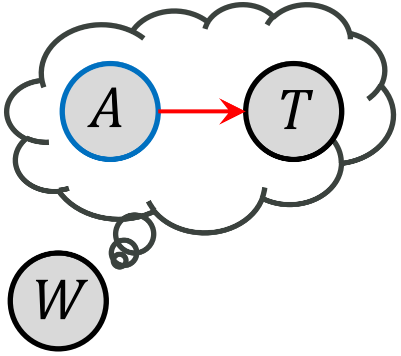
Reality



Causal *Reasoning*

meta SCM (1-variable SCM $\mathcal{M}_2$)

LLM



Causal *Fact*

Observational distribution of $\mathcal{M}_2$ can answer queries w.r.t interventional distribution of $\mathcal{M}_1$

But is this true in LLMs? ***Partially***

Correlation of Causal Facts

Conjecture (Correlation of Causal Facts). Let $Q \subset L_2(\mathcal{M}_1)$ be an interventional query from SCM $\mathcal{M}_1$ and $A \subset L_1(\mathcal{M}_2)$ be its answer from the observational distribution of a meta-SCM $\mathcal{M}_2$. Let f be the LLM's predictive model. Then,

$$f(Q) = A \Leftrightarrow f(Q) \text{ minimizes prediction error}$$

What does this mean? We can answer interventional queries from observational data of meta SCMs iff the predictor is optimal and

1. Causal fact was observed in the training data
2. Causal fact is **highly correlated** with the query

*What about LLM fine-tuning? CCF applies since fine-tuning can be interpreted as finding the right meta SCM for downstream task.

Three Tasks for Evaluation

Do LLMs give causal facts as answers to causal queries?

- Common sense reasoning of LLMs (physics, logic, etc.)
  - Intuitive Physics
  - Propositional Logic (A causes B, B causes C, does A cause C?)
  - Chain of Thought (similar to PL but with phrases)
- Causal Discovery
  - Can LLMs recover causal facts?
- LLMs that use embeddings of knowledge base facts
  - How “causal” are LLM embeddings?

UNIVERSITY OF ARKANSAS

Reasoning Capability

Physics

	Intuitive Physics						
	Rolling (8)	Support (8)	Collisions (4)	Seesaw (4)	Weights (5)	Tools (7)	Accuracy
GPT-3	**6**	**5**	**4**	**2**	**2**	3	**61.11%** 22 / 36
Luminous	1	0	0	1	1	2	11.11% 5 / 36
OPT	2	0	1	0	0	**4**	19.44% 7 / 36
GPT-4	**7**	**8**	**4**	**3**	**5**	**5**	**91.66%** (!) 32 / 36

"There is a tilted board above a bucket. Where does a ball end up if it is placed on the board?"	
GPT-4	The ball will end up in the bucket.
GPT-3	The ball will end up in the bucket.
Luminous	A: The ball will roll down the board. *[continued]*
OPT	A. It will fall into the bucket. B. It will fall into the bucket. *[continued]*

Reasoning Capability

Physics

“ [‘A kilogram of metal is heavier than a kilogram of feathers’] is what most people say, but in reality”	
GPT-4	a kilogram of metal weighs the same as a kilogram of feathers. The weight is the same because a kilogram is a kilogram regardless of the material being weighed. The difference lies in the volume or space that each material occupies. Metal is *[end of output]*
GPT-3	they weigh the same.
Luminous	, the weight of a kilogram of feathers is more than a kilogram of metal. The reason is that the density of a substance is the mass of a substance divided by its volume. *[continued]*
OPT	, a kilogram of feathers is heavier than a kilogram of metal. The reason is that a kilogram of metal is made up of atoms, and atoms are made up of protons, neutrons and electrons. *[repeating]*

Reasoning Capability

Propositional Logic

Ordered: A causes B, B causes C, etc.

Randomized: H causes B, B causes R, etc.

		Causal Chains (Basic Prop. Logic)											
		N=2	3	4	5	6	7	8	9	10	Subchains (4)	Randomized (7)	Accuracy
Zero shot	GPT-3		✓	✓	✓			✓		✓	2	2	45.00% 9/20
	Luminous	✓				✓	✓	✓	✓		1	4	50.00% 10/20
	OPT		✓			✓					0	2	20.00% 4/20
CoT prompting	GPT-3 (CoT 4,6)	✓	✓	✓	✓	✓	✓	✓	✓	✓	**4**	**7**	**100.00%** 20/20
	Luminous (CoT 1)	✓	✓	✓	✓	✓	✓	✓	✓	✓	3	3	75.00% * 15/20
	OPT (CoT 4)	✓	✓	✓	✓	✓	✓	✓	✓	✓	3	4	80.00% * 16/20
	GPT-4	✓	✓	✓	✓	✓	✓	✓	✓	✓	**4**	**7**	**100.00%** (!)

May have been in training data

- GPT-3 can handle shorter chains
- Luminous outperforms in randomized order for zero shot
- CoT prompts help chain reasoning performance (8-shot)

"Q: If Y causes Z and X causes Y. Does X cause Z?
A: Because X causes Y and Y causes Z, X causes Z. The answer is yes."

Reasoning Capability

Propositional Logic (Chain-of-Thought Reasoning)

“If A causes B and B causes C. Does A cause C?”	
GPT-4	Yes, in a sense, A does cause C indirectly through B. This is known as a causal chain. However, the relationship may not be as strong or direct as the relationship between A and B or B and C. Other factors could also influence *[end of output]*
GPT-3	It is possible that A causes C, but it is also possible that A and C are unrelated.
Luminous	A: The answer is no. *[continued]*
OPT	If A causes B and B causes C, then A causes C. If A causes B and B causes C, then A does not cause C. *[continued]*

Simply yes/no

Reasoning Capability

Propositional Logic (Chain-of-Thought Reasoning)

"If A causes B and B causes C. Does A cause B?"	
GPT-4	Yes, as per the initial statement, A causes B.
GPT-3	There is not enough information to answer this question.
Luminous	A: The answer is no. *[continued]*
OPT	If A causes B and B causes C, then A causes C. If A causes B and B does not cause C, then A does not cause C. *[repeating]*

Reasoning Capability

Natural Word Chain

	Natural Word Chain			
	Real World (5)	Imaginary (6)	Mixed (4)	Accuracy
GPT-4	4	**6**	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)	**5**	**6**	**4**	**100.00%**
GPT-3 (CoT 2)	**5**	3	3	73.33%
Luminous (CoT 4)	2	5	2	60.00%
OPT (CoT 1,4)	3, 1	5,**6**	2,3	66.66%

Real words

Made-up words

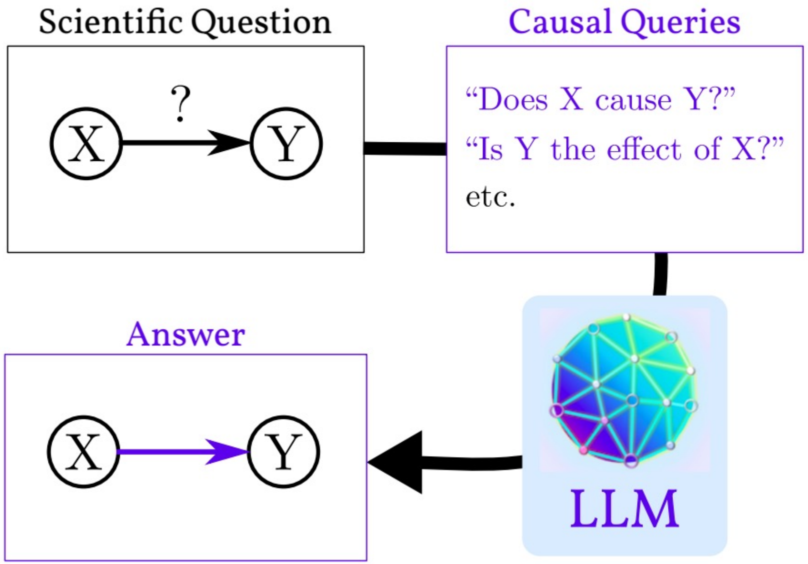
Real and Made-up

Causal Discovery

- What if we want LLMs to recall the right fact?
- Can LLMs recover causal graphs with good prompt?
- Symmetric vs. Asymmetric querying
  - Are X and Y causally **related**?
  - Is there a causal **connection** between X and Y?
  - Is there a **causality** between X and Y?
  - Does X **cause** Y?
  - Does X **influence** Y?
- Number of queries:

  2 * nodes * edges * num_queries

GPT-3 predicts sparser graphs consistent with reality



LLMs increase decisiveness with asymmetric queries

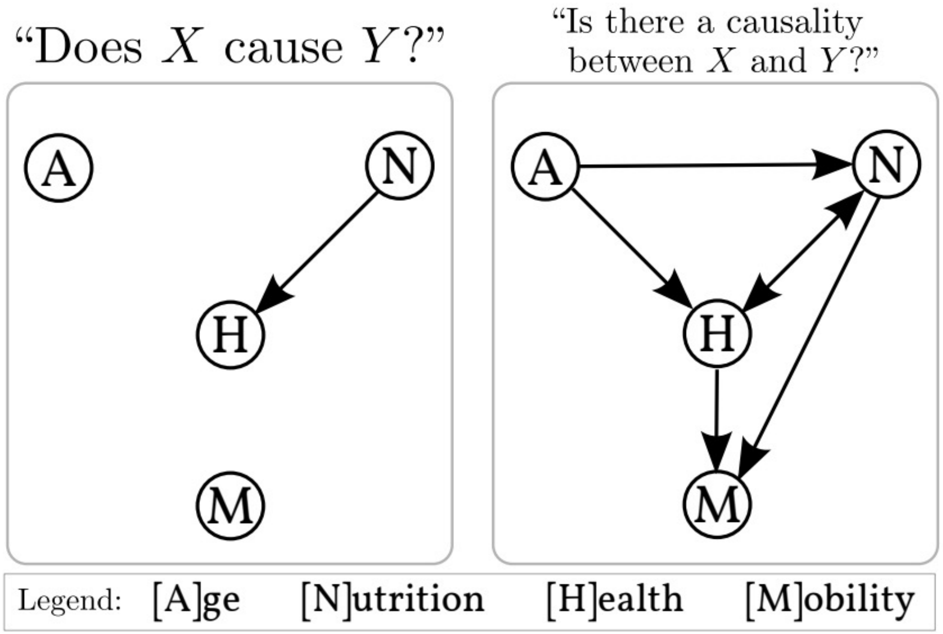
Causal Discovery

	Metric	Altitude	Health	Driving	Recovery	Cancer	Earthquake	LLM
Causal Graph	SID ↓	**0.80**$_{\pm 0.40}$	**7.20**$_{\pm 0.75}$	3.00$_{\pm 0.89}$	**4.00**$_{\pm 1.79}$	11.80$_{\pm 4.66}$	**11.40**$_{\pm 1.50}$	GPT-3
		1.40$_{\pm 0.80}$	9.80$_{\pm 2.99}$	**2.40**$_{\pm 1.20}$	**4.00**$_{\pm 2.53}$	13.20$_{\pm 7.55}$	-	GPT-4
		1.20$_{\pm 0.98}$	10.60$_{\pm 1.85}$	6.00$_{\pm 0.00}$	5.40$_{\pm 1.20}$	**11.40**$_{\pm 3.07}$	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$_{\pm 0.40}$	**4.00**$_{\pm 0.63}$	2.60$_{\pm 0.49}$	**2.20**$_{\pm 0.40}$	**7.00**$_{\pm 1.41}$	**4.60**$_{\pm 0.80}$	GPT-3
		0.80$_{\pm 0.40}$	6.20$_{\pm 2.23}$	**1.60**$_{\pm 0.80}$	2.80$_{\pm 1.60}$	7.40$_{\pm 1.62}$	-	GPT-4
		0.60$_{\pm 0.49}$	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$_{\pm 0.11}$	**0.12**$_{\pm 0.15}$	GPT-3
		0.60$_{\pm 0.33}$	**0.55**$_{\pm 0.06}$	**0.64**$_{\pm 0.10}$	**0.63**$_{\pm 0.19}$	**0.51**$_{\pm 0.04}$	-	GPT-4
		0.80$_{\pm 0.16}$	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	**0.62**	**0.33**	0.50	0.69	0.00	GPT-3
		0.50	0.61	0.17	**0.83**	**0.85**	-	GPT-4
		1.00	0.53	0.17	0.17	0.38	0.26	Luminous
		0.50	0.25	0.25	0.33	0.28	**0.47**	OPT

sparser

directedness

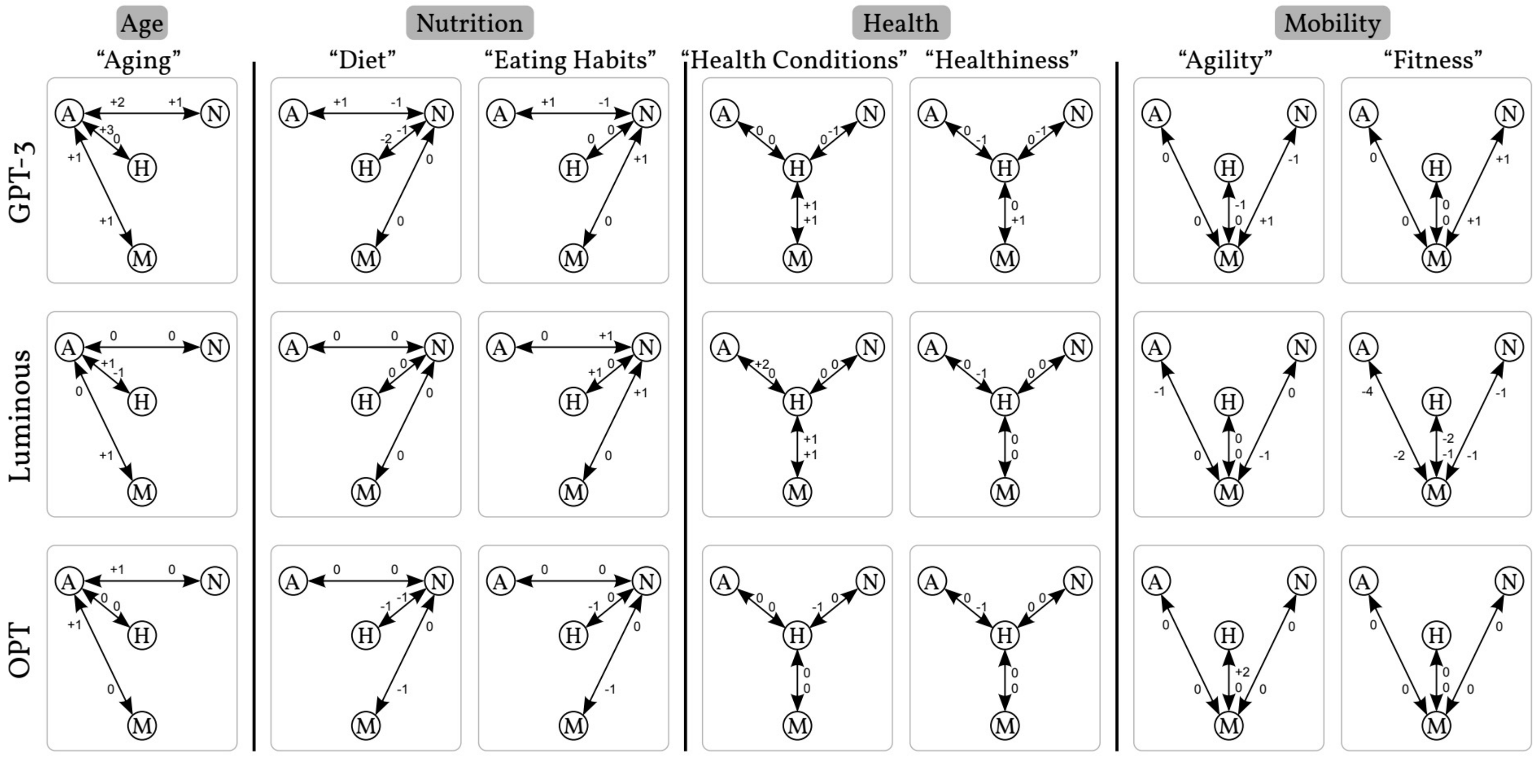
Sensitivity to Wording



Suggests that embedding for "cause" is far away from embedding for "causality" -> enough to get correct answer?

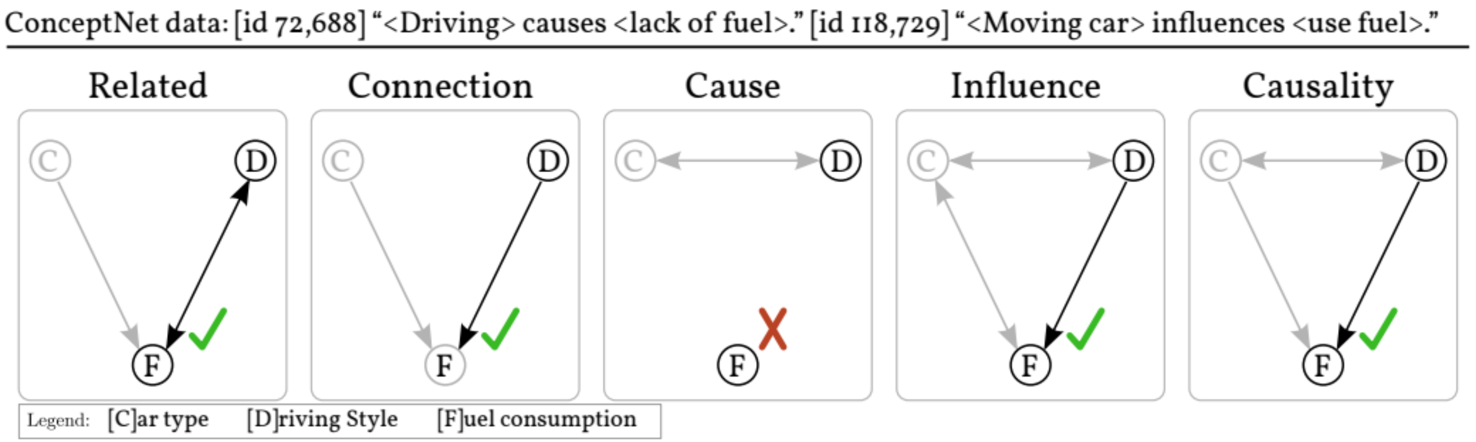
Causal Discovery

Sensitivity to Synonym Swaps



Knowledge-base Embeddings

- Instead of embedding separate words, embed relations from KG
- KG embedding can be “causal”
- Relations from dataset present in ConceptNet KG
- Each wording -> get nearest ConceptNet fact (suggests causal link?)



Conclusion

- LLMs are not causal because
  - no inductive reasoning since no physical measurements
  - violate Pearl's causal hierarchy theorem
- LLMs are causal parrots...but not very good ones
  - LLMs use causal facts found in training data
- Even if LLMs can exploit correlations exposed by meta SCM talking about causal facts, LLM would need infinite amount of meta SCM data to consistently "talk" causal
- We should still try to benchmark causal reasoning capability of LLMs